FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Contrastes de especificación para modelos de distribución

Carlos Rodríguez Ameal

2018/2019

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Contrastes de especificación para modelos de distribución

Carlos Rodríguez Ameal

Julio de 2019

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Coñecemento: Estadística e Investigación Operativa
Título: Contrastes de especificación para modelos de distribución
Breve descrición do contido
Se trata de revisar los contrastes de especificación más notables para modelos de distribución de una variable aleatoria discreta o continua. 1. Contrastes de especificación basados en la estimación de una distribución multinomial 2. Contrastes de especificación basados en la estimación de la función de distribución 3. Contrastes de especificación basados en la estimación de la función de densidad 4. Ilustración en bases de datos simulados

Índice general

Resumen viii

Introducción xi

1. Contrastes basados en la estimación de una distribución multinomial 1

1.1. Introducción: el contraste χ^2 de Pearson 1

1.2. Marco teórico 2

1.2.1. La distribución multinomial 2

1.2.2. Test de razón de verosimilitudes 5

1.2.3. Test de razón e verosimilitudes aplicado a la multinomial 6

1.3. Contrastes de especificación basados en la multinomial 9

1.3.1. Hipótesis simple 9

1.3.2. Hipótesis compuesta 10

1.4. La familia de estadísticos de divergencia 12

1.5. Comentarios sobre el método para distribuciones continuas 14

2. Contrastes basados en la función de distribución empírica 15

2.1. La función de distribución empírica 15

2.2. El proceso empírico 17

2.3. El estadístico de Kolmogorov-Smirnov 19

2.3.1. El estadístico KS para hipótesis compuestas 21

2.3.2. El estadístico KS aplicado al caso discreto 22

2.4. Estadísticos de Anderson-Darling 24

2.4.1. Comentarios finales 25

3. Contrastes basados en la estimación de la función de densidad 27

3.1. Estimación de la función de densidad 28

3.1.1. El histograma 28

3.1.2. Estimación de densidad kernel 29
3.1.3. Análisis del estimador 31
3.2. Contraste de hipótesis 33
3.2.1. Estadísticos de contraste 33
3.2.2. Comportamiento asintótico 35
3.2.3. Hipótesis compuesta 37
3.3. Comentarios finales 37

4. Ilustración en bases de datos simulados **39**
4.1. Hipótesis simple 40
4.1.1. Contraste usando los estadísticos de divergencia 40
4.1.2. Contraste con los estadísticos basados en el proceso empírico 43
4.1.3. Contrastes con los estadísticos basados el la estimación de la función de densidad 46
4.2. Hipótesis compuesta 49

Conclusión **53**

Código de R **57**
.1. Hipótesis simple 57
.1.1. Primer capítulo 57
.1.2. Segundo capítulo 59
.1.3. Tercer capítulo 63
.2. Hipótesis compuesta 65
.2.1. Primer capítulo 65
.2.2. Capítulo 2 68

Bibliografía **75**

Resumen

Este trabajo trata de revisar los fundamentos matemáticos sobre los que se apoyan las tres familias más importantes de contrastes de especificación para una muestra dada: la basada en la estimación de una multinomial, la que se sustenta en la función de distribución empírica y la que utiliza el estimador de densidad kernel; así como de recopilar los estadísticos de mayor importancia que han surgido en cada una de ellas y que dan lugar a los distintos test de uso actual. También se exponen las diferentes pautas para los métodos de contraste, las cuales han sido establecidas por diversos autores tras estudios tanto teóricos como por simulación. Finalmente, se ilustran los test propuestos utilizando bases de datos simuladas para el caso de una hipótesis simple y compuesta.

Abstract

In this dissertation we aim to review the mathematical fundamentals that conform the three main families of goodness-of-fit tests: the one which relies on the estimation of a multinomial distribucition from the sample, the one which is based on the empirical cumulative distribution function and that which employs the kernel density estimation function. We introduce the main statistics and which originate from each perspective, and we give the most important criteria that need to be followed when implementing the different methods according to the suggestions of several authors. Eventually, we ilustrate the tests applying them to simulated samples for both the case of the simple and the composite null hypothesis.

Introducción

A menudo se suele presentar la inferencia estadística como la ciencia de transformar lo concreto en lo abstracto. Dada una muestra finita resultante de un proceso aleatorio, intentamos buscar modelos probabilísticos que se ajusten a ella. De esta forma, podemos obtener información sobre dicho proceso y, en última instancia, ser capaces de hacer predicciones. En este contexto entran los contraste de especificación, también llamados de bondad de ajuste (nombre que quizás sea más esclarecedor, aunque se usa en más ámbitos que el que nos ocupa), y que suponen una de las herramientas más importantes con las que cuenta la inferencia estadística para este fin.

Su implementación está muy extendida en todos los ámbitos de las ciencias naturales y sociales, y de hecho han sido motivados en gran medida por ellas, especialmente por la biología evolutiva y la genética. Además, muchas técnicas estadísticas en un principio independientes, a menudo precisan de su uso en algún punto. Es el caso, por ejemplo, del modelo lineal general, que tiene como hipótesis fundamental la normalidad de los errores y que por tanto necesita de un test que permita contrastarla. Todo ello explica el constante trabajo que se ha puesto en el desarrollo de este tipo de contrastes desde el famoso artículo de Pearson en 1900, y que perdura en la actualidad.

En este trabajo vamos a tratar las tres familias principales de entre todos los contrastes de especificación que existen: la de test basados en la estimación de la multinomial, la de basados en la función de distribución empírica y la de basados en la estimación de la función de densidad. Todas tienen en común el objetivo de proporcionar estadísticos que funcionen como medidas de la discrepancia entre la muestra y nuestra distribución hipotética, y de emplear la teoría de la probabilidad para obtener la distribución exacta de estos estadísticos o, en su defecto, de determinar su distribución asintótica para poder realizar inferencia.

En realidad, existen muchas familias de contrastes y todas comparten esta misma filosofía. El que nos hayamos decidido por estas dos últimas familias en concreto responde al hecho de que, a la hora de medir esa discrepancia, utilicen los elementos más importantes a la hora de caracterizar una variable aleatoria: su función de distribución y su función de densidad. Así, nuestra hipótesis nula inicial, H_0, es que la muestra siga una distribución

determinada, lo cual expresamos como $H_0 : S_n \in X$, siendo n el tamaño de la muestra y X la distribución que sospechamos que puede seguir. La sustituiremos en cada caso por $H_0 : F = F_0$ o $H_0 : f = f_0$, siendo F y f las funciones de distribución y densidad respectivamente de la variable aleatoria de la que verdaderamente proviene nuestra muestra, y F_0 y f_0 las de X. A este contraste, en el que nuestra hipótesis nula solo contempla una única distribución, lo denominamos simple. Es el caso en el que nos preguntamos si S_n viene de una uniforme de intervalo $[0, 1]$, o de una normal de media $\mu = 3$ y varianza $\sigma^2 = 5$.

También puede darse el caso de que en vez de simplemente preguntarnos si nuestra muestra sigue una distribución totalmente especificada, lo que queramos contrastar es si viene de una familia más amplia de distribuciones. Es decir, si proviene de una exponencial para algún λ, o si se ajusta a una normal para algún μ y algún σ. En este caso, nuestra hipótesis nula es compuesta, y se suele expresar como $H_0 : F \in \{F_\theta : \theta \in \Theta\}$ o $H_0 : f \in \{f_\theta : \theta \in \Theta\}$, donde θ es el vector de parámetros que distingue cada función dentro de cada familia.

Por su parte, el caso del primer capítulo es algo especial, pues su importancia radica sobretodo en motivos históricos. Los primeros contrastes de especificación pertenecen a esta familia, y se adaptaron a falta de otras herramientas más potentes a todo tipo de variables, tanto discretas (para las que son óptimos), como continuas (para las que no lo son tanto). Sin embargo, siguen siendo estudiados hoy en día, y su popularidad es indiscutible. En cuanto a sus hipótesis, su formalización se basa en sustituir $H_0 : S_n \in X$ por un contraste paramétrico basado en la multinomial cuya formalización explicaremos detenidamente en el capítulo.

Para terminar, ya que este trabajo trata sobre contrastes, es necesario introducir ciertos conceptos fundamentales y que vamos a estar empleando continuamente. Ya hemos hablado de qué expresión toma nuestra hipótesis nula, que se denota por H_0, y que se corresponde con nuestra suposición, la cual aceptaremos a no ser que tengamos fuertes evidencias de su falsedad. En este otro caso aceptaríamos la hipótesis que se denomina alternativa, H_1, y que simplemente denota: $H_1 : H_0\ es\ falsa$. En este contexto podemos cometer dos clases de errores. El primero ocurre cuando nuestra hipótesis nula es cierta y nosotros, siguiendo el resultado de nuestro test, la rechazamos. Es lo que denominamos error de tipo I. A la inversa, es también posible que nuestra hipótesis nula sea falsa y nosotros la aceptemos, con lo que estaríamos cometiendo el error de tipo II.

Como ya hemos dicho, nuestros estadísticos funcionan como distancias entre nuestra muestra y nuestra distribución hipotética, por lo que los test rechazarán la hipótesis nula cuando estas tomen valores mayores a un determinado punto (cuando «caigan» en la región crítica), y la aceptaremos cuando se queden por debajo (cuando «caigan» en la región

de aceptación). Este punto se denomina punto crítico, y para fijarlo debemos atender a dos cuestiones. La primera es el comportamiento de nuestro estadístico; dado que es una función de una muestra sacada de una variable aleatoria, también será él mismo una variable aleatoria, y por tanto tendrá una distribución que podemos conocer o estimar. La segunda es qué probabilidad, α, de cometer un error de tipo I vayamos a permitir, es decir, el nivel de significación que vayamos a exigir. Nosotros tomaremos siempre $\alpha = 0{,}05$, que uno de los niveles clásicos y más universalmente empleados. De esta forma, fijaremos el punto crítico para que solo una de cada veinte veces nuestro estadístico calculado a partir de una muestra obtenida bajo la hipótesis nula vaya a caer en la región de rechazo. Como es evidente, intentar minimizar el error de tipo I conlleva aumentar la probabilidad de cometer error de tipo II, probabilidad que se suele denotar por β y se denomina potencia.

Una vez que estos conceptos han quedado asentados, para seguir el trabajo correctamente solo debería ser necesaria una cierta base de probabilidad elemental. Esperamos que todo el trabajo que hemos puesto haya servido para crear un texto claro y riguroso, y que hayamos acertado en el enfoque a la hora de tratar las numerosas cuestiones incluidas, muchas de ellas poco relacionadas entre sí, pero que de alguna forma u otra se encuentran en su utilidad dentro del ámbito que nos ocupa: el de los contrastes de especificación.

Capítulo 1

Contrastes basados en la estimación de una distribución multinomial

El más antiguo y quizás mejor conocido de entre todos los contrastes de especificación existentes es el de la χ^2 de Pearson, introducido en su famoso trabajo de 1900. Su relevancia histórica es indiscutible, ya que durante varias décadas de principios del siglo XX fue el único test de bondad de ajuste disponible, empleándose universalmente no solo para el contraste de multinomiales, sino que fue más tarde modificado para distribuciones continuas. Estas eran transformadas en multinomiales dividiendo su conjunto de llegada en distintos intervalos y hallando las probabilidades asociadas a cada uno.

Es evidente que esta forma de proceder supone una pérdida importante de información, y, en general, es más aconsejable utilizar los otros tipos de contrastes que estudiaremos en los próximos capítulos, mejor adaptados a distribuciones continuas. Sin embargo, la importancia de la distribución multinomial, junto con la relevancia histórica y conceptual del método y el hecho de que varios de los tests más modernos aún se basen en él, justifican la inclusión para el estudio de esta familia de contrastes en el presente trabajo.

1.1. Introducción: el contraste χ^2 de Pearson

Supongamos que tenemos una muestra que sospechamos proviene de una variable aleatoria discreta cuya distribución conocemos. Nos referiremos a cada posible valor de la muestra a través de un índice i, y por O_i denotamos el número de veces que hemos obtenido ese resultado. Para estudiar si se cumple nuestra hipótesis, lo natural sería comparar de alguna forma dichas observaciones con los resultados más probables que obtendríamos en el caso de que nuestra suposición sea cierta. A estas cantidades las denotamos E_i.

Esta idea sencilla es la que tiene Pearson cuando propone su famoso estadístico, que

tiene la siguiente expresión:

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \tag{1.1}$$

La intuición de Pearson es bastante clara. El estadístico acumula las diferencias al cuadrado entre las observaciones y los resultados esperados, pero multiplicadas por un factor de ponderación: el inverso de lo esperado. El motivo de la elección de este factor de ponderación es bastante fácil de comprender: no es lo mismo que la diferencia sea de 3 cuando esperamos 4 observaciones que cuando esperamos 40.

Si bien la fórmula del estadístico de Pearson parece bastante directa, lo novedoso de su trabajo fue que en su célebre artículo de 1900 llegó a describir correctamente el comportamiento límite que tomaba: una χ^2. Aunque el descubrimiento en sí de esta distribución no se le concede, su trabajo supuso un punto crucial en la estadística matemática, que en su época se encontraba estancada en la cuestión de hasta qué punto las distribuciones estándares, concretamente la normal, servían para modelizar universalmente los diferentes procesos aleatorios. Alejándose de esa postura, llega al test de la χ^2, dando lugar al primer contraste de especificación propiamente dicho.

Pearson no llegó a hablar de distribución multinomial, pero esta está implícita en su comprensión de los datos «esperables» frente a los observados. Nosotros, desde un enfoque más moderno, utilizaremos las propiedades de la multinomial para demostrar la convergencia asintótica del estadístico de Pearson a una χ^2. Además, daremos un enfoque alternativo basándonos en el test de razón de verosimilitudes, popularizado décadas más tarde por Fisher. Justamente, este último debe su interés por la bioestadística (la genética mendeliana y la biología evolutiva fueron dos campos que motivaron numerosos avances en las técnicas estadísticas de principios del siglo XX) a la lectura de las obras de Pearson, y ambos protagonizaron una sonada polémica al discrepar acerca del comportamiento del X^2 cuando se usa para contrastar hipótesis compuestas. Estas cuestiones teóricas las introduciremos a continuación.

1.2. Marco teórico

1.2.1. La distribución multinomial

Supongamos que tenemos una variable aleatoria X que puede tomar k resultados diferentes, de forma que cada suceso i tiene una probabilidad de ocurrir p_i. Cada uno de sus resultados obtenidos viene modelizado por una variable categórica, mientras que el vector que mide las frecuencias observadas de los resultados obtenidos en n intentos se denomina

multinomial. Así, una variable categórica supone una generalización de una Bernouilli, y una multinomial de una binomial, cuando pasamos de 2 a k posibles resultados. Al igual que podemos definir la distribución binomial como una suma de variables Bernouilli independientes, la multinomial puede expresarse como una suma de variables categóricas i.i.d.[1] Más formalmente:

Definición 1.1. Definimos la variable categórica asociada al suceso X_r de una variable aleatoria X, $r \in \{1, 2, ..., n\}$, como el vector:

$$\xi_r = (I_{\{X_r=1\}}, I_{\{X_r=2\}}, ..., I_{\{X_r=k\}}), \tag{1.2}$$

donde $I_{\{X_r=i\}}$ es la función indicatriz asociada al resultado i de la variable aleatoria X_r.

Por tanto, el vector toma el valor 1 en la posición i-ésima y 0 en todas las demás con probabilidad $p_i = P(X_r = i)$. La suma de n variables categóricas resulta en una multinomial, que expresamos como sigue:

Definición 1.2. Definimos la multinomial de tamaño k y con parámetros n y $\mathbf{p} = (p_1, p_2, ..., p_n)$ tal que $\sum_{i=1}^{n} p_i = 1$, como el vector aleatorio $\mathbf{N} = (N_1, N_2, ..., N_k)$ con función de probabilidad:

$$P\{N_1 = n_1, N_2 = n_2, ..., N_k = n_k\} = \frac{n!}{n_1!n_2!...n_k!} p^{n_1} p^{n_2} ... p^{n_k} \tag{1.3}$$

Una forma alternativa de definir ambas variables consiste en suprimir la última coordenada de los vectores (que pasarán así a ser $k-1$ dimensionales), de forma que si el resultado aleatorio r toma un valor en la k-ésima categoría, la variable categórica correspondiente tenga todas las coordenadas iguales a 0. En cuanto a la variable multinomial, si queremos ver cuántos resultados han ocurrido en la categoría k tras n intentos, no tenemos más que hallar la resta $n - \sum_{i=1}^{k-1} n_i$. Ambas representaciones son equivalentes y las utilizaremos indistintamente según nos convenga.

Ahora enunciaremos las propiedades de la distribución categórica cuya demostración es inmediata:

Lema 1.3. *Sea $\xi_r = (I_{\{X_r=1\}}, I_{\{X_r=2\}}, ..., I_{\{X_r=k\}})$ una variable categórica con parámetro $\boldsymbol{p} = (p_1, p_2, ..., p_n)$. Se verifica:*

1. $E(\xi_r) = \boldsymbol{p}$
2. *Su matriz de covarianzas $\Sigma = (\sigma)_{ij}$ cumple:*

[1] Independientes e idénticamente distribuidas

$$\sigma_{ij} = E[I_{\{X_r=i\}} \cdot I_{\{X_r=j\}}] - p_i p_j = \begin{cases} -p_i p_j & si\ i \neq j \\ p_i(1-p_i) & si\ i = j \end{cases}$$

A partir de la expresión de la multinomial como suma de variables categóricas i.i.d. $\mathbf{N} = \sum_{i=1}^{n} \xi_i$ y de las propiedades de la esperanza y la covarianza podemos probar el siguiente lema:

Lema 1.4. *Sea $\boldsymbol{N} = (N_1, N_2, ..., N_k)$ una multinomial de parámetros n y $\boldsymbol{p} = (p_1, p_2, ..., p_n)$. Se verifica:*

1. $N_i \in Binomial(n; p_i)$ *para* $i = 1, 2, ..., k$.
2. $E(N_i) = np_i$ *para* $i = 1, 2, ..., k$
3. $Var(N_i) = np_i(1-p_i)$ *para* $i = 1, 2, ..., k$
4. $Cov(N_i, N_j) = -np_i p_j$

Para terminar, presentaremos un resultado sobre la convergencia de la multinomial que nos será útil más adelante para estudiar el comportamiento asintótico nulo del estadístico de Pearson. Dado que tenemos la restricción: $\sum_{i=1}^{k} N_i = n$, la matriz de covarianzas del vector tiene rango $n-1$. Para solventar este contratiempo, emplearemos la definición alternativa de la multinomial bajo la forma de un vector de dimensión $k-1$.

Teorema 1.5. *Sea $\xi_r = (I_{\{X_r=1\}}, I_{\{X_r=2\}}, ..., I_{\{X_r=k-1\}})$ la variable categórica asociada al resultado X_r y sean $\boldsymbol{p}$ y Σ su esperanza y matriz de covarianzas respectivamente. Si $\boldsymbol{N} = (N_1, N_2, ..., N_{k-1}) = \sum_{r=1}^{n} \xi_r$ es el vector multinomial que acumula n resultados, se cumple cuando $n \to \infty$ que:*

$$\frac{\boldsymbol{N} - n\boldsymbol{p}}{\sqrt{n}} \Sigma^{-1/2} \xrightarrow{d} N_{k-1}(0, I).^{2}$$

,

Demostración. Definamos $\eta_r = (\xi_r - \mathbf{p})\Sigma^{-1/2}$, que tiene media igual a $\mathbf{0}$ y matriz de covarianzas I_{k-1} . Entonces, por el Teorema Central del Límite:

$$\frac{\boldsymbol{N} - n\boldsymbol{p}}{\sqrt{n}} \Sigma^{-1/2} = \frac{1}{\sqrt{n}} \sum_{r=1}^{n} \eta_r \xrightarrow{d} N_{k-1}(0, I)$$

□

[2]Por $X_n \xrightarrow{d} X$ denotamos la convergencia de la sucesión de variabes X_n a la variable X en distribución.

1.2.2. Test de razón de verosimilitudes

Al principio de este capítulo hemos introducido el estadístico de Pearson, nacido de su intuición matemática aplicada al problema de cómo medir la discrepancia entre una muestra y una distribución esperada. Pearson fue capaz de llegar a la distribución límite bajo la nula de este estadístico particular usando sus conocimientos de probabilidad, y de crear uno de los primeros test de contraste: el de la χ^2. Sin embargo, a pesar de la enorme importancia que su trabajo tuvo en su época y de la influencia que sigue teniendo hoy en día, si el mismo problema se plantease a un estadístico actual, probablemente la primera forma de abordar el contraste que se le ocurriese sería utilizando uno de los métodos más universales: el test de razón de verosimilitudes. Lo interesante es que cuando este método fue creado y estudiado, sirvió para confirmar el resultado al que Pearson había llegado tres décadas antes.

El test se aplica siempre que tengamos una muestra que sigue una distribución F_θ con fución de densidad f_θ, especificada excepto por un parámetro θ que varía en un espacio paramétrico $\Theta \subset \mathbb{R}^k$, y queremos contrastar $H_0 : \theta \in \Theta_0$ frente a $H_1 : \theta \in \Theta \setminus \Theta_0$ donde $\Theta_0 \subset \Theta$. El método se basa en el concepto de verosimilitud de la muestra: $f_\theta(x_1, x_2, ..., x_n)$, que funciona como una medida de lo bien que «explica» θ los resultados obtenidos. Así, $\sup_{\theta \in \Theta_0} f_\theta(x_1, ..., x_n)$ supone un índice de «la mejor explicación» de la muestra bajo H_0, y el cual vamos a querer comparar con $\sup_{\theta \in \Theta} f_\theta(x_1, ..., x_n)$, que se corresponder con la mejor explicación posible que existe entre todos los valores posibles del parámetro. Los estimadores máximo verosímil serán los parámetros asociados a estos valores, respectivamente: $\hat{\theta}_0$ y $\hat{\theta}$ tales que $f_{\hat{\theta}_0}(x_1, ..., x_n) = \sup_{\theta \in \Theta_0} f_\theta(x_1, ..., x_n)$ y $f_{\hat{\theta}}(x_1, ..., x_n) = \sup_{\theta \in \Theta} f_\theta(x_1, ..., x_n)$. Para comparar ambas cantidades calculamos el cociente entre ellas, que es lo que denominamos rezón de verosimilitudes:

$$\Lambda(x_1, ..., x_n) = \frac{\sup_{\theta \in \Theta_0} f_\theta(x_1, ..., x_n)}{\sup_{\theta \in \Theta} f_\theta(x_1, ..., x_n)} = \frac{f_{\hat{\theta}_0}(x_1, ..., x_n)}{f_{\hat{\theta}}(x_1, ..., x_n)} \tag{1.4}$$

Un test de razón de verosimilitudes rechazará la hipótesis nula cuando $\Lambda(x_1, ..., x_n) < c$ para un c que fijaremos en función del nivel de significación que busquemos. Para ello podemos intentar hallar la distribución exacta de Λ, que en ocasiones puede ser expresado como un estadístico más simple. Sin embargo, esto es muy difícil en general, por lo que es necesario contar con algún resultado que nos dé información sobre su comportamiento asintótico. Más precisamente, usaremos el fenómeno de Wilks, que caracteriza la convergencia del logritmo de este cociente.

Teorema 1.6. *Supongamos que la hipótesis nula dependa de q parámetros, es decir:* $\Theta_0 = \{\theta \in \Theta : \theta_i = g_i(w_1, ..., w_q)$ *para* $i = 1, ..., k$; *con* $(w_1, ..., w_q) \in \Omega\}$ *siendo* Ω *un abierto de* $\mathbb{R}^q$ *y* g_i *funciones con derivadas parciales de orden* 1 *continuas. Bajo ciertas hipótesis de regularidad (se puede consultar el manual Principios de Inferencia Estadística de Ricardo Vélez para más información), y cuando* $n \to \infty$, *si la hipótesis nula es cierta se cumple que :*

$$-2\ log\Lambda(x_1, ..., x_n) \xrightarrow{d} \chi^2_{k-q}$$

Es decir, conforme la muestra se hace más grande, el estadístico $-2log\Lambda(x_1, ..., x_n)$ converge a una χ^2 cuyos grados de libertad se corresponden a la diferencia de las dimensiones entre Θ y Θ_0. No hemos explicitado las hipótesis de regularidad del manual que estamos siguiendo (el de Ricardo Vélez) puesto que el propio autor afirma que estas son suficientes pero no necesarias, y que la validez del resultado se suele aceptar sin reparos en los contextos usuales.

Gracias a este resultado, ya podemos definir nuestro test, que rechazará la hipótesis nula con un nivel de significación α siempre que: $-2\ log\Lambda > \chi^2_{k-1;\alpha} (\Leftrightarrow \Lambda < e^{\frac{-\chi^2_{k-1;\alpha}}{2}})$

1.2.3. Test de razón e verosimilitudes aplicado a la multinomial

Ahora nos interesa aplicar la teoría de la sección anterior al caso particular de una multinomial. Supongamos que nuestra muestra $x_1, ..., x_n$ proviene de una variable discreta que toma k valores diferentes y tiene a $\mathbf{N}$ como vector multinomial con parámetro $\mathbf{p}$, y que queremos contrastar $H_0 : \mathbf{p} = \mathbf{p}_0$ frente a $H_1 : \mathbf{p} \neq \mathbf{p}_0$. Usando la notación que hemos introducido antes, tenemos que Θ y Θ_0 tienen dimensión $k - 1$ ($\mathbf{p} \in \mathbb{R}^k$ con la restricción $\sum_{i=1}^k p_i = 1$) y 0 respectivamente.

En primer lugar, para calcular la razón de verosimilitudes, veamos que el estimador máximo verosímil del parámetro p_i es $\hat{p}_i = N_i/n$. De aquí en adelante denotaremos de la misma manera a cada una de las componentes del vector aleatorio N_i, como al número de observaciones de la muestra iguales a i (es decir, no distinguiremos entre N_i y n_i). Entonces, el problema de hallar el estimador máximo verosímil se corresponde con hallar los parámetros $\mathbf{p}$ de $\mathbf{N}$ que hacen más probable los N_i, es decir, que maximizan la función de probabilidad:

$$\hat{\mathbf{p}} = ArgMax_{\mathbf{p} \in \mathbb{R}^k, \sum_{i=1}^k p_i = 1, p_i \geq 0} \frac{n!}{N_1!N_2!...N_k!} p_1^{N_1} p_2^{N_2} ...(1 - p_1 - p_2 - ... - p_{k-1})^{N_k}$$

Aplicando logaritmo a la función de probabilidad (el logaritmo es creciente e inyectivo) e igualando las derivadas parciales respecto de los p_i a cero para buscar los argumentos máximos, obtenemos las ecuaciones de verosimilitud:

$$\begin{cases} \frac{N_j}{p_j} - \frac{N_k}{1-p_1-p_2-...-p_{k-1}} = 0 \\ j = 1, 2, ..., k-1 \end{cases}$$

por lo que la solución verifica:

$$\frac{n_k}{1-\hat{p}_1-\hat{p}_2-...-\hat{p}_{k-1}} = \frac{n_1}{\hat{p}_1} = \frac{n_2}{\hat{p}_2} = ... = \frac{n_{k-1}}{\hat{p}_{k-1}}$$

De un sencillo cálculo obtenemos la relación $\hat{p}_i = N_i/n$. La matriz de derivadas segundas es definida positiva, con lo que es un máximo relativo, y como en la frontera alguna de las probabilidades se anula, así lo hace la función de densidad y concluimos que el máximo es global. Ahora podemos obtener la razón de verosimilitudes:

$$\Lambda(N_1, N_2, ..., N_k) = \frac{(p_1^0)^{N_1}(p_2^0)^{N_2}...(p_k^0)^{N_k}}{\hat{p}_1^{N_1}\hat{p}_2^{N_2}...\hat{p}_k^{N_k}} = \prod_{i=1}^{k} \left(\frac{p_i^0}{\hat{p}_i}\right)^{N_i} \tag{1.5}$$

Podemos definir entonces un región crítica: $\{\Lambda < c\} = \{-2\ \log\Lambda > h\}$, que es la región de valores para los cuales rechazamos la hipótesis nula. Para poder elegir valor k en función del nivel de significación que queramos nos basamos en el estadístico:

$$G^2 = -2\ \log\Lambda = 2\sum_{i=1}^{k} N_i(\log\ \hat{p}_i - \log\ p_i^0), \tag{1.6}$$

que tiene una distribución asintótica χ^2_{k-1} suponiendo la hipótesis nula $H_0 : p = p^0$ como consecuencia del teorema 1.6.

Es decir, este estadístico se comporta de misma forma que el ideado por Pearson en el contexto de los contrastes de especificación. Recordemos que este último se definía como la suma de las diferencias al cuadrado entre los valores observados y los esperados divididas por los esperados. La formalización de la multinomial nos aporta una nueva notación para este estadístico más acorde con el tono del trabajo:

$$X^2 = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0} \tag{1.7}$$

Es sorprendente que, a pesar de nacer en dos contextos totalmente diferentes, ambos estadísticos son muy parecidos en la práctica. La clave está en la intuición de Pearson al elegir como factores de ponderación los $1/np_i^0$, pues admitiendo que $(p_i^0 - \hat{p}_i)/\hat{p}_i$ sean pequeños:

$$log\frac{p_i^0}{\hat{p}_i} = log\left(1 + \frac{p_i^0-\hat{p}_i}{\hat{p}_i}\right) \simeq \frac{p_i^0-\hat{p}_i}{\hat{p}_i} - \frac{1}{2}\left(\frac{p_i^0-\hat{p}_i}{\hat{p}_i}\right)^2$$

y por tanto:

$$log\Lambda \simeq \sum_{i=1}^{k} N_i \frac{p_i^0 - \hat{p}_i}{\hat{p}_i} - \frac{1}{2}\sum_{i=1}^{k} \frac{(p_i^0 - \hat{p}_i)^2}{\hat{p}_i^2} = n\sum_{i=1}^{k}(P_i^0 - \hat{p}_i) - \frac{1}{2}\sum_{i=1}^{k}\frac{n^2}{N_i}(\hat{p}_i - p_i^0)^2 = \\ - \frac{1}{2}\sum_{i=1}^{k}\frac{(N_i - np_i^0)^2}{N_i}$$

de forma que:

$$-2log\Lambda \simeq \sum_{i=1}^{k}\frac{(N_i - np_i^0)^2}{N_i} \simeq \sum_{i=1}^{k}\frac{(N_i - np_i^0)^2}{np_i^0} \tag{1.8}$$

puesto que $N_i \simeq np_i^0$ para $i = 1, ..., k$.

Este similitud provoca, como ya hemos dicho, un mismo comportamiento asintótico para ambos estadísticos. Para G^2 este resultado se prueba echando mano de la teoría del test de razón de verosimilitudes. Pero es fácil de ver para el estadístico de Pearson con unos pocos cálculos. Ello da lugar al siguiente teorema:

Teorema 1.7. *Cuando $n \to \infty$ y si la hipótesis nula es cierta:*

$$X^2 = \sum_{i=1}^{k}\frac{(N_i - np_i^0)^2}{np_i^0} \xrightarrow{d} \chi^2_{k-1}$$

Demostración. Por el teorema 1.5 y simplemente usando la definición de la χ^2 tenemos:

$$\left(\frac{\boldsymbol{N} - n\boldsymbol{p}}{\sqrt{n}}\right)\Sigma^{-1}\left(\frac{\boldsymbol{N} - n\boldsymbol{p}}{\sqrt{n}}\right) \xrightarrow{d} \chi^2_{k-1}$$

Pero calculando la matriz Σ^{-1} y desarrollando esta producto (consúltese el manual de Ricardo Vélez para más detalles) se llega precisamente a la igualdad:

$$\left(\frac{\boldsymbol{N} - n\boldsymbol{p}}{\sqrt{n}}\right)\Sigma^{-1}\left(\frac{\boldsymbol{N} - n\boldsymbol{p}}{\sqrt{n}}\right) = X^2$$

□

Contamos entonces con dos test que rechazarán la hipótesis nula con un nivel de significación α cuando su estadístico correspondiente tome un valor mayor que $\chi^2_{1-k;\alpha}$. La precisión de estos test depende en buena medida de cómo de buena sea la aproximación de la χ^2 para las distribuciones exactas de los estadísticos. El criterio más extendido que se suele pedir es que los valores esperados cumplan $np_i^0 > 5$, que se suele combinar con restricciones sobre el número de obervaciones (normalmente se pedirá $n > 30$) y el de intervalos (cuando no imposibilite la primera pauta se cogerá $k > 5$).

1.3. Contrastes de especificación basados en la multinomial

1.3.1. Hipótesis simple

Supongamos que tenemos n datos de una muestra aleatoria simple que sigue una distribución desconocida F, y que deseamos contrastar la hipótesis nula $H_0 : F = F_0$, donde F_0 es una función completamente especificada (no depende de ningún parámetro de valor desconocido). De esta forma la hipótesis nula es simple, y su alternativa $H_1 : F \neq F_0$ está compuesta por todas las distribuciones distintas a F_0.

Recordamos que este es el caso más sencillo, en el que en vez de contrastar si la muestra ha sido obtenida de una familia de distribuciones se hace para una F_0 de la cuál conocemos su función de distribución, y por tanto la probabilidad p_i^0 asociada a cualquier intervalo A_i en el que tome valores. Dividiendo entonces el recorrido en A_1, A_2,...,A_k intervalos disjuntos, inmediatamente podemos considerar el vector aleatorio $\mathbf{N} = (N_1,N_2,...,N_k)$, con N_i igual al número de observaciones de la muestra en cada subconjunto A_i, y que sigue una distribución multinomial con cada parámetro p_i igual a la probabilidad de que F dé un valor en A_i. La idea consiste en sustituir la hipótesis inicial por $H_0 : p = p^0$ frente a $H_1 : p \neq p^0$, siendo $p^0 = (p_1^0, p_2^0, ..., p_k^0)$ el vector con las probabilidades de que F_0 caiga en cada intervalo, y dando lugar al contraste paramétrico ya visto, y que podemos realizar usando los estadísticos X^2 o G^2. Más tarde presentaremos una familia general de estadísticos de contraste y haremos ciertos comentarios sobre sus propiedades.

Está claro que esta forma de proceder tiene una desventaja fundamental. Cuando la distribución hipotética que queremos contrastar es discreta, los intervalos pasan a ser puntos y el contraste de especificación se corresponde desde un primer momento con el de la multinomial que hemos estudiado en la sección anterior. Sin embargo, cuando nuestra F_0 es continua, el método se basa en una discretización de su recorrido, es decir, la transformamos en una multinomial con toda la pérdida de información que eso conlleva. Podemos dar infinitos ejemplos de funciones con distribuciones notadamente diferentes que dan lugar a multinomiales idénticas si los intervalos se eligen de la manera adecuada (pero lo cual no es nada adecuado para nuestro propósito). Así, aceptar la hipótesis nula del contraste multinomial no garantiza en la realidad que podamos aceptar la hipótesis nula del contraste original. La solución a este defecto pasa por elegir el número de intervalos y la posición de sus fronteras de la forma adecuada. Al final del capítulo comentaremos con detenimiento el proceso del contraste aplicado a distribuciones continuas y intentaremos dar ciertas pautas a seguir.

1.3.2. Hipótesis compuesta

Supongamos ahora que en vez querer contrastar la hipótesis nula $H_0 : F = F_0$ para F_0 una distribución completamente especificada, lo que queremos es saber si F pertenece a una familia de distribuciones de las cual conocemos la expresión general de la función de densidad, pero esta es dependiente de uno o varios parámetros no especificados.

Intuitivamente, podríamos intentar solucionar el problema calculando los estimadores máximo verosímiles y realizando el contraste simple con la distribución ya especificada de la familia que nos interesa contrastar. Por ejemplo, si sospechamos que nuestra F es una Normal, calcularíamos la media y la varianza más verosímiles con respecto a la muestra, $\hat{\mu}$ y $\hat{\sigma}$, y que nos determina una distribución de la que conocemos la función de densidad. Podríamos obtener ahora las probabilidades de los k intervalos en los que dividimos $\mathbb{R}$ ($p_i^0 = \int_{A_i} f(x|\hat{\mu}, \hat{\sigma}^2)\ dx$) y usar el estadístico de Pearson para realizar el contraste. Sin embargo, esto es demasiado precipitado, pues actuando de esta manera las p_i^0 dependerían de $\hat{\mu}$ y $\hat{\sigma}$, y por tanto de la muestra, con lo que la demostración del teorema 1.7 deja de ser válida y ya no podemos afirmar que el estadístico X^2 siga convergiendo en distribución a una χ^2_{k-1}.

Una forma de evitar este contratiempo sería utilizar dos muestras diferentes de la misma distribución, una para hallar los estimadores de máxima verosimilitud y especificar F_0, y la otra para realizar el contraste de hipótesis nula simple: $F = F_0$. Sin embargo veremos que existen métodos más adecuados en los que la muestra contribuye tanto en la estimación de los parámetros como en la obtención del estadístico de contraste. Para ello dividimos como antes el conjunto de posibles valores en $A_1, A_2, ..., A_k$ intervalos disjuntos y consideramos la variable multinomial $\boldsymbol{N} = (N_1, N_2, ..., N_k)$ de tamaño n y vector de probabilidades $\boldsymbol{p} = (p_1, p_2, ..., p_k)$. La hipótesis nula incial es $H_0 : F \in \{F_\theta : \theta \in \Theta \subset \mathbb{R}^q\}$, que sutituimos por el contraste de los parámetros de $\boldsymbol{N}$ y reformulamos de forma análoga al caso simple: $H_0 : \boldsymbol{p} = \boldsymbol{\pi}(\theta)$. Así obtenemos la región de probabilidades $\mathcal{R} = \{\boldsymbol{\pi}(\theta) : \theta \in \Theta\}$. Consideraremos siempre que $q < k - 1$.

Como al estimar θ por máxima verosimilitud estamos seleccionando la distribución de la familia que mejor se adapta a la muestra, podemos razonar que tendremos que ser más exigentes con las frecuencias obtenidas para aceptar la hipótesis nula que en el caso de la hipótesis es simple. La forma de cuantificar esta idea puede ser empleando el test de razón de verosimilitudes, que nos permite construir con seguridad un estadístico que mida la discrepancia entre los datos y los valores esperados, y que converja en distribución a una χ^2 cuyos grados de libertad sabemos determinar.

El estimador máximo verosímil de θ interesa para realizar el contraste entre multinomiales, por lo que se busca el parámetro que hace más probable la muestra obtenida pero

en la versión discretizada de la distribución, con lo que sustituimos cada dato por su vector categórico. Así, para la muestra $(x_1, x_2, ..., x_n)$ simplificada en $(\xi_1, \xi_2, ..., \xi_n)$, la función de probabilidad con la que se trabaja es: $f_\theta(\xi_1, \xi_2, ..., \xi_n) = \prod_{i=1}^{k} \pi(\theta)_i^{N_i}$. Suponiendo que las π_i son derivables respecto de θ obtenemos las ecuaciones de verosimilitud:

$$\begin{cases} \displaystyle\sum_{i=1}^{k} \frac{N_i}{\pi_i(\theta)} \frac{\partial}{\partial \theta_j} \pi_i(\theta) = 0 \\ j = 1, 2, ..., q \end{cases} \tag{1.9}$$

que habitualmente tiene una única solución.

Una vez que hemos calculado $\hat{\theta}$ y puesto que el máximo sobre $\boldsymbol{p}$ se sigue alcanzando en $\hat{\boldsymbol{p}}$, ya podemos obtener el estadístico del test:

$$\hat{G}^2 = -2\ log\Lambda = 2 \sum_{i=1}^{k} N_i (log\ \hat{p}_i - log\ \pi_i(\hat{\theta}))$$

que por el teorema 1.6 tiene como distribución asintótica una χ^2_{k-1-q}. Es decir, cada parámetro que debemos estimar de la distribución poblacional disminuye en 1 el número de grados de libertad de la distribución asintótica de la razón de verosimilitudes.

Por razones históricas, el test de razón de verosimilitudes no es muy frecuentemente empleado para hacer estos contrastes de especificación. Sin embargo, nos ha servido para resolver el problema de los grados de libertad de la χ^2 cuando hay varios parámetros desconocidos. Precisamente esta cuestión dio lugar a una sonada controversia entre el Fisher y Pearson: el primero consideraba (correctamente) que en el caso de la hipótesis compuesta $H_0 : F \in \{F_\theta : \theta \in \Theta \subset \mathbb{R}^q\}$, el nuevo estadístico $\hat{X}^2 = \sum_{i=1}^{k} \frac{(N_i - n\pi_i(\hat{\theta}))^2}{n\pi_i(\hat{\theta})}$ convergía a una χ^2_{k-1-q}, mientras que el segundo defendía que el número de grados de libertad seguía siendo $k-1$. Fisher consiguió probar que efectivamente había que tener en cuenta el número de parámetros estimados, y por ello $\hat{X}^2$ es conocido como el estadístico de Pearson-Fisher.

El siguiente teorema (Principios de inferencia estadística UNED 2012) permite afirmar la convergencia de $\hat{X}^2$:

Teorema 1.8. *Sea Θ un abierto de $\mathbb{R}^q$ $(q < k-1)$, en el cual:*
i)Existen y son continuas $\frac{\partial}{\partial \theta_j}\pi_i(\theta)$ para cada $i = 1, ..., k$ y $j = 1, ..., q$
ii)La matriz $k \times q$ de términos $\frac{1}{\sqrt{\pi_i(\theta)}} \frac{\partial}{\partial \theta_j}\pi_i(\theta)$ tiene rango q.
Entonces dada una sucesión $\{\hat{\theta}_n\}_{n\in\mathbb{N}}$ de soluciones del sistema de ecuaciones 1,9*, tal que $\hat{\theta}_n \to \theta$, se verifica que:*

$$\hat{X}^2 = \sum_{i=1}^{k} \frac{(N_i - n\pi_i(\hat{\theta}_n))^2}{n\pi_i(\hat{\theta}_n)} \xrightarrow{d} \chi^2_{k-1-q}$$

Este resultado es de hecho válido para sucesiones de estimadores más generales que simplemente las soluciones del sistema de ecuaciones 1,9. De hecho, podemos afirmarlo para todo estimador *BAN* de θ.

Definición 1.9. Decimos que un estimador $\hat{\theta}$ es *BAN* (*best asymptotically normal*) si (1) es consistente[3], (2) su distribución asintótica es normal y (3) es asintóticamente eficiente[4].

Todo estimador por máxima verosimilitud es *BAN*, sin embargo esta propiedad no es una equivalencia. Un contraejemplo es $\hat{\theta} = ArgMin_{\theta\in\Theta} X_n^2(\theta)$, que también es un estimador *BAN*. La idea es clara, suponiendo que X^2 es una medición de la discrepancia entre los datos observados con los esperados, parece una buena idea tomar como estimador de los parámetros a aquel valor que acerque lo máximo posible la distribución hipotética a la realidad. Este razonamiento se puede aplicar también tomando el estadístico G^2 como medidor de discrepancia. Obtendríamos en este caso el estimador máximo verosímil. En la próxima sección generalizaremos los estadísticos de contraste, y, por esta dualidad, obtendremos una familia de estimadores *BAN*.

1.4. La familia de estadísticos de divergencia

Hasta ahora hemos presentado los dos estadísticos más importantes para efectuar contrastes de bondad de ajuste: el X^2 y el G^2, y hemos visto que, bajo ciertas condiciones de regularidad y siempre suponiendo cierta la hipótesis nula, tienen el mismo comportamiento asintótico. Otros estadísticos que han sido propuestos son el de Freeman-Tukey:

$$\widehat{F}^2 = 4\sum_{i=1}^{k}\left(\sqrt{N_i} - \sqrt{n\pi_i(\hat{\theta})}\right)^2,$$

el del logaritmo de la razón de verosimilitud modificado:

$$\widehat{GM}^2 = 2\sum_{i=1}^{k} n\pi_i(\hat{\theta})(log\ n\pi_i(\hat{\theta}) - log\ N_i),$$

o el X^2 modificado por Neyman:

$$\widehat{X}^2 = \sum_{i=1}^{k}\frac{(N_i - n\pi_i(\hat{\theta}))^2}{N_i}$$

[3]Un estimador de un parámetro β es consistente si converge a β en probabilidad

[4]Es decir, su distribución asintótica tiene la menor varianza de entre todos los estimadores que cumplen (1) y (2).

Se ha probado por varios autores que todos estos estadísticos tienen la misma distribución asintótica χ^2 que X^2 y G^2, bajo las condiciones ya vistas. Este comportamiento similar no es anecdótico, sino que fue estudiado por Cressie y Read (1984), los cuales llegaron a una familia de estadísticos caracterizados por un único parámetro: el *power-divergence statistic*.

Definición 1.10. El *power-divergence statistic* se define como:

$$2nI^{\lambda}(N;\pi(\hat{\theta})) = \frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{k} N_i\left[\left(\frac{N_i}{n\pi_i(\hat{\theta})}\right)^{\lambda} - 1\right], \tag{1.10}$$

donde $\lambda \in (-\infty, +\infty)$ es el parámetro de la familia.

Esta fórmula no está definida para $\lambda \in \{-1, 0\}$; pero estas discontinuidades son evitables y podemos extender el estadístico de forma continua a $\mathbb{R}$. Para ello solo hay que emplear el infinitésimo $lim_{h\to 0}(t^h - 1)/h = log(t)$ y obtenemos los estadísticos $\widehat{G}^2$ para $\lambda = 0$ y $\widehat{GM}^2$ para $\lambda = -1$. Por su parte, si evaluamos 1.10 con $\lambda = 1, -1/2$, y -2 obtenemos los estadísticos X^2, F^2 y NM^2 respectivamente. El siguiente teorema nos permite afirmar que el comportamiento asintótico de cualquier estimador de esta familia es el mismo que el de X^2 y G^2 bajo las condiciones ya vistas:

Teorema 1.11. *Supongamos que H_0 es cierta y que $\hat{\theta}$ es un estimador BAN de θ. Entones cuando $n \to \infty$,*

$$2nI^{\lambda}(\boldsymbol{N};\hat{\theta}) \xrightarrow{d} \chi^2_{k-1-q} \qquad \infty < \lambda < \infty$$

Ahora podemos generalizar el concepto de estimador máximo verosímil a una nueva familia de estimadores que reducen el valor de $2nI^{\lambda}(N;\pi(\hat{\theta}))$, cada uno para un λ. Es decir:

$$\hat{\theta}_{\lambda} = ArgMin_{\theta\in\Theta} I^{\lambda}(\mathbf{N};\theta)$$

Además, Read y Cressie (1988) probaron que esta es una familia de estimadores *BAN*, siempre y cuando $\pi(\theta)$ satisfaga ciertas condiciones de regularidad. Nótese que esto implica la convergencia de $2nI^{\lambda}(\mathbf{N};\hat{\theta}_{\lambda'})$, aún cuando $\lambda \neq \lambda'$, como es el caso del estadístico de Pearson ($\lambda = 1$), que utilizábamos con el estimador máximo verosímil ($\lambda' = 0$).

Ahora es natural preguntarse qué estadístico tiene mejores propiedades, en el sentido de que su convergencia hacia la χ^2 es más rápida. El problema cobra importancia sobretodo para muestras pequeñas, para las cuales Cressie y Read concluyeron que de hecho el estadístico de Pearson X^2 tiene un buen comportamiento. En general sugieren tomar siempre $\lambda \geq 0$, por lo que recomiendan no emplear el test de verosimilitudes, que estaría en el límite de esta región de estadísticos "buenos". De entre ellos, el estadístico que parece tener un mejor comportamiento parece ser el de parámetro $\lambda = 2/3$.

1.5. Comentarios sobre el método para distribuciones continuas

Obviamente, los estadísticos vistos hasta ahora son óptimos para contrastar si una muestra proviene de una multinomial, pero pierden fuerza al ser aplicados para el contraste de distribuciones continuas. Históricamente este hecho era obviado por la falta de métodos mejores, pero hoy en día la existencia de contrastes como los que se verán en los dos próximos capítulos hace injustificada la utilización de este procedimiento, que al final depende de la simplificación de la distribución hipotética y de los datos de la muestra. Por ello solo haremos unos comentarios poco profundos sobre las cuestiones que quedan por tratar.

Empezando con el caso de la hipótesis simple, no se ha explicado cómo se deben elegir el número de intervalos (k) ni sus fronteras. Lo ideal sería poder hacerlo de modo que el test ganara fuerza contra la alternativa que nos interese más, pero en general no vamos a tener una idea muy clara de cuál es la alternativa más factible a la muestra obtenida. Varios autores han sugerido que las fronteras de los intervalos sean tomadas de forma que sean equiprobables bajo la hipótesis nula, de esta forma Cressie y Read vieron que la χ^2 supone una buena aproximación para frecuencias esperadas tan bajas como $1/4$ (con $n \geq 10$ y $k \geq 3$). Por otro lado, en general parece tener buen resultado el aumentar k conforme tengamos muestras más grandes como es bastante intuitivo: cuantos más intervalos tomemos más información de la muestra es retenida y por tanto mayor poder tendrá el test. Pero dar una regla para el ratio de k a partir de n es complicado: si el número de intervalos aumenta ya no se puede asegurar que lo haga el número de frecuencias esperables, con lo que toda la teoría asintótica de la que depende el comportamiento de los estadísticos deja de ser válida.

Cuando trabajamos con una hipótesis compuesta el problema se vuelve aún más complejo. Podemos tener la tentación de estimar θ directamente de la muestra para evitar la pérdida de información que supone el agrupar las observaciones, pero entonces los estimadores ya no tienen el comportamiento asintótico descrito y se vuelven más complicados. Por otro lado, la cuestión de cómo elegir los intervalos deviene un reto difícil. El criterio de equiprobabilidad deja de suponer una simplificación en absoluto, ya que la probabilidad de cualquier intervalo que tomemos dependerá de la estimación del parámetro y esta lo hace de los datos, con lo que la teoría coge una gran complejidad. La solución que se suele emplear por su sencillez y porque asegura la validez de los resultados visto en el capítulo, consiste en tomar los intervalos de forma puramente aleatoria. abandonando la regla de la equiprobabilidad.

Capítulo 2

Contrastes basados en la función de distribución empírica

Un enfoque diferente al basado en la multinomial para los contrastes de bondad de ajuste es el desarrollado incialmente por la escuela rusa y alemana de probabilidad. Interesado por realizar un contraste específicamente adaptado a las variables continuas (para las cuales el test de la χ^2 de Pearson presenta serias limitaciones), Kolmogorov tiene la idea de medir su discrepancia respecto de la hipótesis a partir de lo que caracteriza una función: su función de distribución. Así, dada una muesta $x_1, x_2, ..., x_n$ obtenida de una misma variable aleatoria, para contrastar la hipótesis nula: $H_0 : F = F_0$ frente a $H_1 : F \neq F_0$, calcula un estimador de la función de distribución que sigue la muestra y lo compara con la F_0. Este estimador será la función de distribución empírica, cuya definición y propiedades constituyen la base de todos los estadísticos de este capítulo.

2.1. La función de distribución empírica

La función de distribución empírica (que denominaremos FDE de ahora en adelante), es un estimador muy intuitivo de la función de distribución F de una variable aleatoria X. Dada una muestra de n elementos, si consideramos la interpretación probabilística de F, $F(x) = Pr(t \leq x)$, la FDE «aproxima» esta probabilidad calculando la proporción de observaciones de la muestra menores o iguales que x. Bien formalizado, tenemos la siguiente definición:

Definición 2.1. Dada $S_n = x_1, x_2, ..., x_n$ una muestra de n observaciones i.i.d, y supongamos sin pérdida de generalidad que $x_1 \leq x_2 \leq ... \leq x_n$. Definimos:

$$\hat{F}_n(x) = \frac{1}{n}\#\{X_i \in S_n : X_i \leq x\} = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \quad [1]$$

Es inmediato ver que $\hat{F}_n$ es una función no decreciente y escalonada, con cada escalón extendiéndose entre x_{i-1} y x_i y $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$. También cumple que es continua por la derecha y tiene límite por la izquierda de todo punto x, con lo que en particular $\hat{F}_n$ es una función de distribución.

Ejemplo de función de distribución empírica

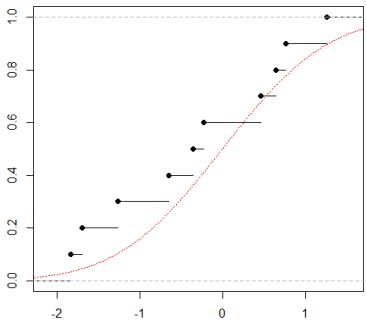


Figura 2.1: Los segmentos negros se corresponden con los escalones de la función de distribución empírica obtenida a partir de 20 datos simulados de una normal estándar. Su función de distribución real se representa con una línea roja punteada

Ahora nos centraremos en el comportamiento puntual de $\hat{F}_n$. Dado que $n\hat{F}_n(x)$ es la función que cuenta el número de observaciones de la muestra S_n menores o iguales que x, y que cada dato tiene una probabilidad $F(x)$ de ser menor o igual que x, es obvio que $n\hat{F}_n(x)$ sigue una Binomial de parámetros n y $F(x)$ (de hecho al definirla la hemos expresado como la suma de variables Bernouilli). De aquí deducimos las siguientes propiedades:

Lema 2.2. *1.* $E(\hat{F}_n(x)) = F(x)\ \forall x, \forall n$

2. Por la ley fuerte de los grandes números, cuando $n \to \infty$

$$\hat{F}_n(x) \xrightarrow{a.s.} F(x)\ \forall x \quad [2]$$

3. Por el Teorema Central de Límite, cuando $n \to \infty$

$$\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{d} N(0, F(x)(1-F(x)))\ \forall x$$

[1] $I(X_i \leq x)$ es el indicador del evento $(X_i \leq x)$, que fijada x se comporta como una Bernouilli con $p = F(x)$

[2] a.s. significa *almost sure* en inglés y se usa para denotar que la convergencia es casi segura.

Como corolario inmediato de la propiedad (2), podemos afirmar que la FDE es un estimador consistente de $F(x)$ en cada punto x.

Estas propiedades nos dan información sobre el comportamiento asintótico de $\hat{F}_n(x)$, pero hacen referencia a convergencias puntuales. Sin embargo, la propiedad (2) puede ser extendida para la convergencia uniforme de $\hat{F}_n(x)$ a $F(x)$. Esto es lo que afirma el Teorema de Givenko-Cantelli:

Teorema 2.3. *Givenko-Cantelli. Con la notación anterior, y cuando* $n \to \infty$,

$$\sup_x |(\hat{F}_n(x) - F(x))| \xrightarrow{a.s.} 0$$

Estos resultados aseguran que $\hat{F}_n$ se acerca cada vez más a F cuando la muestra es grande. Por ello, si queremos realizar el contraste $H_0 : F(x) = F_0(x)$, tiene sentido medir la diferencia entre $\hat{F}_n(x)$ y $F_0(x)$. Esta es justamente la idea subyacente de los test que veremos en esta sección, cuyos estadísticos serán de la forma:

$$T_n = c(n)d(\hat{F}_n, F_0), \tag{2.1}$$

siendo $c(n)$ un factor de escala y $d(.,.)$ una distancia o función de divergencia[3] entre funciones. Nótese que $\hat{F}_n$ es una función de x al igual que F_0, por lo que tiene sentido medir esta diferencia.

Por definición, se cumple que $d(F, F_0) = 0$[4] $\Leftrightarrow H_0$ es cierta. Además, notemos que tal y como $\hat{F}_n$ es un estimador de de F, $d(\hat{F}_n, F_0)$ es un estimador de $d(F, F_0)$, con lo que siempre trabajaremos con medidas cuyos estimadores sean consistentes. Esto implica que cuando n se hace grande, $d(\hat{F}_n, F_0)$ bajo la nula tiende a 0 en probabilidad. Una vez que elegimos d podemos estudiar las propiedades de T_n, para lo que suele ser crucial el comportamiento asintótico de $\sqrt{n}(\hat{F}_n(x) - F(x))$. Sin embargo, los resultados de convergencia puntual no suelen ser suficientes para la mayor parte de los T_n.

2.2. El proceso empírico

A pesar de que la FDE constituye la base sobre la cual se articula la familia de test que estudiamos en este capítulo, suele ser más conveniente trabajar directamente con el

[3] Una función de divergencia es más débil que una distancia en el sentido de que no es necesariamente simétrica ni cumple la desigualdad triangular.

[4] Esta igualdad puede ser entendida como la igualdad estricta de funciones (que podría ser contrastada con el estadístico de Kolmogorov-Smirnov), o, más en general, como la igualdad de las clases de F y F_0 en el espacio de funciones L_n (a la que se limita los estadísticos de Anderson-Darling). Como estamos en un contexto probabilístico nos llega con que se cumpla la igualdad para casi todo punto, con lo que entenderemos de esta última forma la igualdad de la hipótesis nula y no nos preocuparemos más por esta cuestión

proceso empírico, que se definirá a continuación.

Definición 2.4. Dada una muestra de n observaciones de una variable aleatoria X con función de distribución $F(x)$, definimos el proceso empírico como la función aleatoria $B_n(x) = \sqrt{n}(\hat{F}_n(x) - F(x))$.

Por el lema 2.8, conocemos el comportamiento asintótico de B_n para cada x. Pero esto no será suficiente, y precisamos estudiar el comportamiento asintótico funcional del proceso empírico. Un primer paso para enfrentarnos a este problema es considerar el vector k-dimensional, donde cada coordenada corresponde a B_n evaluada en diferentes puntos del dominio de F, y estudiar sus propiedades. Ello da lugar al siguiente resultado:

Proposición 2.5. *Para todo $x_1, x_2, ..., x_k$ valores en el dominio de $F(x)$, cuando $n \to \infty$ se tiene:*

$$(B_n(x_1), ..., B_n(x_k)) \xrightarrow{d} (B(x_1), ..., B(x_k)) \in N(0_k, \Sigma),$$

donde $\Sigma = (\sigma_{ij})$, con $\sigma_{ij} = Cov\{B(x_i), B(x_j)\} = F(x_i \wedge x_j) - F(x_i)F(x_j)$ [5]

Demostración. Es consecuencia directa del Teorema Central del Límite Multivariante, solo hay que tener en cuenta que $B_n(x) = \sqrt{n}(\hat{F}_n(x) - E[I(X_i \leq x)])$ y que $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$. Veamos que $Cov\{B(x_l), B(x_m)\} = F(x_l \wedge x_m) - F(x_l)F(x_m)$.

Por el TCLM:

$$Cov\{B(x_l), B(x_m)\} = Cov\{I(X_1 \leq x_l), I(X_1 \leq x_m)\} =$$
$$E[I(X_1 \leq x_l) \cdot I(X_1 \leq x_m)] - E[I(X_1 \leq x_l)] \cdot E[I(X_1 \leq x_m)].$$

Usando que $I(X_1 \leq x_l) \cdot I(X_1 \leq x_m) = I(X_1 \leq min\{x_l, x_m\})$, obtenemos el resultado. □

Conforme tomamos más puntos y aumentamos la dimensión del vector, este se vuelve una mejor aproximación de la función B_n. Pero para llegar a un TCL funcional no vale solo con dejar que k tiena a infinito, sino que hacen falta más condiciones. Sin embargo, para los resultado de este trabajo será suficiente con pensar un TCL funcional como el límite de un TCL multivariante. Decimos entonces que el proceso empírico B_n converge débilmente al proceso límite B, y lo denotamos por:

$$B_n \xrightarrow{w} B$$

La función límite B es lo que se conoce como proceso Gaussiano, es decir, un conjunto de variables aleatorias indexadas (en este caso por los x), tales que toda colección finita

[5] $x \wedge y = min\{x, y\}$

de dichas variables tiene una distribución normal multivariante. En particular, se trata de un proceso Gaussiano de media cero y covarianza la dada por la proposición 2.5.

Para finalizar esta sección presentamos un último teorema de gran utilidad: el teorema de Mann-Wald:

Teorema 2.6. *Sea g una función continua. Si $B_n \xrightarrow{w} B$, entonces $g(B_n) \xrightarrow{w} g(B)$ cuando $n \to \infty$.*

2.3. El estadístico de Kolmogorov-Smirnov

En las anteriores secciones hemos recopilado los resultados más importantes referentes a la FDE y el proceso empírico. Ahora podemos introducir ya el primer estadístico de este capítulo y uno de los primeros que se emplearon para realizar un contraste de bondad de ajuste: el de Kolmogorov-Smirnov, a menudo abreviado como KS. Para contrastar la hipótesis nula $H_0 : F = F_0$ frente $H_1 : F \neq F_0$, el estadístico toma la forma:

$$D_n = \sqrt{n} \cdot \sup_x |\hat{F}_n(x) - F_0(x)|, \tag{2.2}$$

con lo que D_n es de la forma de la expresión 2.1, con $d(\hat{F}_n, F_0) = \sup_x |\hat{F}_n(x) - F_0(x)|$ y $c(n) = \sqrt{n}$.

Además, bajo la hipótesis nula $D_n = \sup_x |B_n(x)|$ y así podremos aplicar los resultados de la sección anterior. Nótese que D_n es directamente proporcional al supremo de la diferencia entre la función de la hipótesis G y la FDE. A menudo esta diferencia también se suele escribir como $\hat{F}_n(G^{-1}(p)) - p$ $(p = G(x))$. En muchos manuales el estadístico se presenta sin el factor $\sqrt{n}$, aunque nosotros preferimos nuestra notación por poner de relieve su relación clara con el proceso empírico.

Smirnov introdujo dos estadísticos íntimamente relacionados con D_n, y que se definen como sigue:

$$D_n^+ = \sqrt{n} \cdot \sup_x (\hat{F}_n(x) - F_0(x)) = \sup_x (B_n(x)) \tag{2.3}$$

$$D_n^- = \sqrt{n} \cdot \sup_x (F_0(x) - \hat{F}_n(x)) = \sup_x (-B_n(x)) \tag{2.4}$$

Está claro que D_n^- mide la mayor desviación positiva entre la FDE y la F_0, mientras que D_n^- mide la mayor desviación negativa. Así, se emplean para los contrastes de hipótesis nula $H_0 : F = F_0$ y alternativas $H_1 : F > GF_0$ y $H_1 : F < F_0$ respectivamente, siendo $>$

la relación de orden estocástico[6] . Justamente, combinando ambos estadísticos, obtenemos KS: $D_n = max\{D_n^+, D_n^-\}$.

Normalmente, encontrar el supremo de una función no diferenciable requiere la evaluación de la función en muchos puntos. Sin embargo, si F es una función continua, podemos suponer que: $x_1 < x_2 < ... < x_n$. Así, dado que $\hat{F}_n$ es una función escalonada y G monótona creciente, las expresiones (2.3) y (2.4) trivialmente se simplifican a:

$$D_n^+ = \sqrt{n} \cdot \max_{1 \leq i \leq n} (\frac{i}{n} - F_0(X_{(i)})) \tag{2.5}$$

$$D_n^- = \sqrt{n} \cdot \max_{1 \leq i \leq n} (F_0(X_{(i)}) - \frac{i-1}{n})) \tag{2.6}$$

Para hallar D_n solo hace falta evaluar F_0 en los n valores de la muestra. Además, esta expresión de los estadísticos permite entrever una propiedad muy importante sobre su comportamiento: su distribución nula no depende de F_0.

Proposición 2.7. *Dada una muestra $x_1, x_2, ..., x_n$ de una variable aleatoria X con una función de distribución F continua, las distribuciones nulas de los estadísticos D_n, D_n^+, D_n^- no dependen de F.*

Demostración. Por las expresiones 2.5 y 2.6, si se cumple $H_0 : F = F_0$ deducimos que D_n^+ y D_n^- solo dependen de las variables $U_{(1)} = F(X_{(1)}), U_{(2)} = F(X_{(2)}), ..., U_{(n)} = F(X_{(n)})$, que son los estadísticos de orden de las variables aleatorias $U_1 = F(X_1), U_2 = F(X_2), ..., U_n = F(X_n)$. Pero si F es continua, $U_i = F(X_i)$ sigue una uniforme de intervalo $[0, 1]$ independientemente de F con lo que tenemos el resultado. Escribiendo $D_n = max\{D_n^+, D_n^-\} = \max_{1 \leq i \leq n} (\frac{i}{n} - F(X_{(i)}), F(X_{(i)}) - \frac{i-1}{n})$ tenemos el resultado también para D_n. □

Ahora, si queremos emplear el estadístico KS para inferencia necesitamos conocer su distribución bajo la nula. Como se acaba de ver que esta es libre, este problema suele ser abordado suponiendo que $F(x)$ es una uniforme en $(0, 1)$ y a partir de ahí se halla la distribución de cada D_n. Aunque estas se han computado hasta un n grande, el proceso es bastante tedioso y suele ser más común emplear la distribución asintótica de los D_n. Gracias a los resultados vistos en las dos primeras secciones de este capítulo podemos hallarla muy fácilmente:

Lema 2.8. *Cuando $n \to \infty$ y si $H_0 : F = F_0$ es cierta, $D_n = \sup_x |B_n(x)| \xrightarrow{d} \sup_x |B(x)| =: D$, donde B es el proceso límite de B_n.*

Demostración. Considerando el proceso límite $B_n \xrightarrow{w} B$ es consecuencia directa del teorema 2.6. □

[6] Dadas X e Y dos variables aleatorias con funciones de distribución F y G respectivamente, decimos que X es estocásticamente menor que Y y representamos por $F > G$, si para todo $\forall z$, $Pr\{X < z\} < Pr\{Y < z\}$.

Existe incluso una expresión analítica de la función de distribución de D que se ha comprobado suficientemente buena para $n > 35$, aunque la demostración no se incluirá en este trabajo:

$$F_D(d) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j+1}exp(-2j^2d^2) \tag{2.7}$$

Entonces, si $n > 35$ el test rechazaría la hipótesis nula $H_0 : F = F_0$ con un nivel de significación α, si y solo si $D_n > F_D^{-1}(1-\alpha)$. Como F_D viene expresado en serie de potencias, habría que aproximar el valor $F_D(1-\alpha)$ para realizar el contraste.

2.3.1. El estadístico KS para hipótesis compuestas

Hasta ahora hemos estudiado el estadístico KS en el contexto de los test simples. Sin embargo, análogamente a los estadísticos basados en la multinomial, este puede ser adaptado para realizar también contrastes de hipótesis compuestas. Aunque veremos que no hay una forma general de proceder para ello.

Supongamos que la distribución de la hipótesis, F_0, es conocida hasta un vector de parámetros $\beta_k^t = (\beta_1, ..., \beta_k)$ de la que es dependiente. Denotamos entonces $F_0(x) = F_0(x;\beta)$ para explicitar dicha dependencia. La manera de proceder será sustituir la hipótesis inicial por $H_0 : F(x) = F_0(x;\hat{\beta}_n)$, siendo $\hat{\beta}_n$ un estimador del que se precisará que cumpla varias condiciones más adelante.

Todos los estadísticos de este capítulo están basados en el proceso empírico, con lo que es necesario estudiar cómo se comporta cuando la distribución F_0 no está completamente especificada. Denotamos $B_n(x;\beta) = \sqrt{n}(\hat{F}_n(x) - F_0(x;\beta))$ y $B(x;\beta)$ su proceso límite bajo la hipótesis nula. Como sabemos, este tiene media cero y su función de covarianza cumple: $Cov\{B(x;\beta), B(y;\beta)\} = F_0(x \wedge y;\beta) - F_0(x;\beta)F_0(y;\beta)$. Notemos que a pesar de escribir la función de distribución $F_0(x)$ de la forma $F_0(x;\beta)$, simplemente es una cuestión de notación y la teoría de la sección anterior se aplica normalmente.

El problema aparece cuando β es desconocido y ha de ser sustituido a la hora de calcular los procesos empíricos por su estimador $\hat{\beta}_n$, con lo que naturalmente los resultados vistos hasta ahora dejan de ser válidos. El siguiente teorema muestra la nueva distribución asintótica de $\hat{B}_n(x;\hat{\beta}_n) = \sqrt{n}(\hat{F}_n(x) - F_0(x;\hat{\beta}_n))$, para un tipo muy concreto de estimadores:

Teorema 2.9. *Si $\hat{\beta}_n$ es un estimador lineal localmente asintótico*[7]*, bajo la hipótesis nula el*

[7]Se dice del estimador $\hat{\beta}_n$ de β que verifica la siguiente expresión: $\hat{\beta}_n - \beta = \frac{1}{n}\sum_{i=1}^{n}\psi(X_i;\beta) + o_p(n^{-1/2})$, donde $\psi^t = (\psi_1, ..., \psi_p)$ es una función vectorial continuamente diferenciable de R^p en R^p con media cero y $E\{\psi(X;\beta)\psi^t(X;\beta)\}$ es finito y no singular.

proceso empírico estimado $\hat{B}_n(x)$ *converge débilmente a un proceso Gaussiano* $\hat{B}$ *de media cero y función de covarianza:* $Cov\{\hat{B}(x), \hat{B}(y)\} =$

$$= F_0(x \wedge y; \beta) - F_0(x; \beta)F_0(y; \beta) - \psi^t(x; \beta)h(y; \beta) - \psi^t(y; \beta)h(x; \beta) + h^t(x; \beta)\Sigma_\psi h(y; \beta),$$

donde $h(x; \beta) = \partial F_0(x; \beta)/\partial\beta$, $\Psi(x; \beta) = \int_{-\infty}^{x} \psi(z; \beta)dF_0(z; \beta)$ *y* $\Sigma_\psi = Var(\psi(x; \beta))$.

Hay dos consecuencias muy importantes de la convergencia de $\hat{B}_n$ a $\hat{B}$. La primera es que nos permite hallar la distribución nula del estadístico KS. Así, bajo H_0, cuando $n \to \infty$:

$$\hat{D}_n = D_n(\hat{\beta}_n) = \sqrt{n} \cdot \sup_x |\hat{F}_n(x) - F_0(x; \hat{\beta}_n)| = \sup_x |\hat{B}_n| \xrightarrow{d} \sup_x |\hat{B}(x)| \tag{2.8}$$

La segunda es que la distribución límite depende del β desconocido, además de la distribución F_0. Por tanto, ya no hay una forma general de realizar el contraste. Sin embargo, si F_0 es una distribución invariante a cambios de escala y posición,[8] $\hat{D}$ se simplifica de forma que deja de depender de β (si bien continúa dependiendo de F_0).

Un caso particular de este tipo de distribuciones es la normal, que ha sido estudiado por Lilliefors. Este propuso aplicar el estadístico KS a las variables estandarizadas $Z_i = \frac{X_i - \overline{X}}{\sqrt{S}}$ (siendo $\overline{X}$ y S la media y la cuasivarianza muestral). Para ello calcularíamos FDE tras la estandarización, F_n, y el estadísticos toma la expresión: $\hat{D}_n = \sqrt{n} \cdot \sup_z |F_n(z) - \phi(z)|$, donde ϕ es la distribución de una normal estándar. Lilliefors fue el primero en tabular las distribuciones exactas de los $\hat{D}_n$ y luego se han empleado diferentes métodos computacionales para hallar su distribucón asintótica. Los puntos críticos del estadístico para los p-valores más empleados se pueden consultar en el manual *Nonparametric Statistical Inference* de Jean Dickinson Gibbons y Shubhabrata Chakabort.

2.3.2. El estadístico KS aplicado al caso discreto

Hasta ahora, todos los resultados expuestos sobre las propiedades del estadístico KS se han basado en que su distribución no depende de la de la variable aleatoria de la muestra, y por tanto requieren como hipótesis la continuidad de la función G. Sin embargo, la teoría presentada en las dos primeras secciones, incluyendo el teorema de Givenko-Cantelli 2.3, no precisan de esta suposición, con lo que existe una motivación para emplear el estadístico en el caso de G discreta. También puede interesar agrupar la muestra en k clases, pero ahora querer emplear el KS y no los estadísticos del primer capítulo. Como esto es básicamente una «discretización» de la función de distribución F, ambos problemas son equivalentes.

[8] Se dice que una distribución es invariante a transformaciones de escala y posición si admite función de densidad f_0 y esta cumple: $g(x; \mu, \sigma) = g((x - \mu)/\sigma; 0, 1)$

Empleando la notación del primer capítulo, supongamos que tenemos k clases, cada una con una probabilidad esperada p_i^0 dada por la función F_0, y N_i representa el número de observaciones de cada clase de la muestra. Entonces tenemos que, dado $j \in \{1, ..., k\}$:

$$\hat{F}_n(x_j) = \sum_{i=1}^{j} N_i/n, \qquad G(x_j) = \sum_{i=1}^{j} p_i^0, \tag{2.9}$$

con lo que el estadístico KS toma la forma:

$$D_n = \max_{i \leq j \leq k} \left| \sum_{i=1}^{j} \frac{N_i - np_i^0}{\sqrt{n}} \right| = \max_{i \leq j \leq k} \left| \sum_{i=1}^{j} \frac{O_i - E_i}{\sqrt{n}} \right| \tag{2.10}$$

De esta forma, la distribución de D_n depende de cómo fueron establecidas las clases a la hora de agrupar los datos. La distribución exacta de $\sqrt{n}D_n$ ha sido tabulada (ver *Pettitt and Stephens*). Es especialmente significativa la similitud entre el KS para el caso discreto y el estadístico de Pearson X^2. Justamente, para terminar esta sección, haremos una pequeña comparación entre ambos estadísticos.

Una primera diferencia obvia es que el estadístico de Pearson requiere que los datos estén agrupados en clases, al contrario que el KS. Por tanto, para distribuciones continuas el KS tiene un uso más completo de la muestra, influyendo cada dato de forma propia, y su uso supone no enfrentarse al problema del la elección del número de clases y de las fronteras de cada una. Además, la distribución exacta de los D_n es conocida y ha sido tabulada para todo n, mientras que a la hora de trabajar con los estadísticos de divergencia solo podemos basarnos en sus distribuciones asintóticas a la χ^2, que supone una buena aproximación siempre y cuando tengamos suficientes datos y las frecuencias esperadas sean lo bastante grandes. Sin embargo, los estadísticos de divergencia tienen la ventaja de tener una distribución asintótica conocida cuando hay presentes parámetros desconocidos (recordamos que cada parámetro resta un grado de libertad de la χ^2 límite), mientras que $\hat{D}_n$ tiene una distribución diferente a D_n, para la cual no contamos con una fórmula general.

Para finalizar, cabe decir que el test del KS tiene mayor poder que el de la chi-cuadrado de Pearson, tanto para datos categorizados como para distribuciones continuas. Con lo que se puede concluir que, exceptuando los casos de hipótesis compuestas para las que no se conozca la distribución límite del $\hat{D}_n$, el estadístico KS es el más adecuado de los dos.

2.4. Estadísticos de Anderson-Darling

Hasta ahora hemos estudiado con detenimiento el estadístico KS y su comportamiento a la hora de contrastar hipótesis simples o compuestas, y para distribuciones continuas y discretas. Sin embargo, este solo es un ejemplo del conjunto de estadísticos construibles siguiendo la fórmula 2.1. Otra familia muy importante de estadísticos es la introducida por Anderson-Darling:

$$T_n = \int_S w(G(x))B_n^2(x)dG(x), \tag{2.11}$$

donde $w(.)$ actúa como una función de peso.

Cuando $w(u) = 1$ para todo $0 \leq u \leq 1$, el estadístico suele ser conocido como el estadístico de Cramér-von Mises, que abreviaremos por CvM. Además, aunque hemos presentado una colección de estadísticos indexada por una función de peso, en la práctica solo hay un único estadístico de Anderson-Darling que es corrientemente empleado (aparte del CvM): el que tiene $w(x) = 1/(u(1-u))$ como función de peso, y que por tanto será al que nos refiramos cuando hablemos de estadístico de Anderson-Darling y abreviaremos por AD. La razón por la cual se emplea esa función de peso particular es que tiene un efecto estabilizador en la varianza: $\sqrt{w(u)}B_n(u)$ tiene varianza constante igual a 1.

A pesar de que calcular el valor de los estadísticos pueda parecer complicado, para CvM y AD existen expresiones explícitas simples. Denotémolos de forma respectiva por W_n y A_n para una muestra de tamaño n. Sean $U_i = G(X_i)$ y $U_{(i)}$ la estadística de orden i-ésimo de las variables $U_1, ..., U_n$. Entonces:

$$A_n = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)(log\ U_{(i)} + log(1 - U_{(n+1-i)})) \tag{2.12}$$

$$W_n = \frac{1}{12n} + \sum_{i=1}^{n}\left(U_{(i)} - \frac{2i-1}{2n}\right)^2 \tag{2.13}$$

Cuando el contraste involucra una hipótesis compuesta: $H_0 : F(x) = G(x;\beta)$ se procede de manera similar al test KS. Primero, calcularíamos el estimador de β, $\hat{\beta}_n$, que suponemos asintóticamente lineal, y sustituimos $G(x;\beta)$ por $G(x;\hat{\beta}_n)$.

Menos es conocido sobre el comportamiento exacto y asintótico de los estadísticos CvM y AD que sobre el del KS. Para todo estadístico de Anderson-Darling T_n ($\hat{T}_n$ cuando trabajamos con hipótesis compuestas), los resultados vistos en la sección anterior siguen siendo válidos, a saber: la distribución de T_n bajo la nula no depende de G y usando

de nuevo la convergencia débil del proceso empírico B_n podemos extender el lema 2.8 al siguiente teorema:

Teorema 2.10. *Si* $\int_0^1 w(t)B^2(t)dt < \infty$ *y* $\hat{\beta}_n$ *es un estimador asintóticamente lineal de* β*, entonces, bajo la hipótesis nula simple y compuesta respectivamente:*

$$T_n \xrightarrow{d} \int_0^1 w(t)B^2(t)dt \tag{2.14}$$

$$\hat{T}_n \xrightarrow{d} \int_0^1 w(t)\hat{B}^2(t)dt, \tag{2.15}$$

cuando $n \to \infty$.

Ninguna de estas fórmulas es verdaderamente útil para hallar directamente los puntos críticos de los estadísticos. En el caso de la hipótesis compuesta recordemos que $\hat{B}$ depende de G y de β, con lo que en general no se pueden tabular los p-valores. Aunque sí que es posible hacerlo cuando G es invariante a cambios de locación y escala, como en el caso de que G sea una normal o una exponencial.

En el caso de la hipótesis simple se han podido obtener varios resultados interesantes. A partir de sus funciones características, Anderson y Darling hallaron expresiones de las distribuciones asintótitcas de W_n y A_n, W y A respectivamente, como sumas infinitas de variables aleatorias:

$$W = \sum_{j=1}^{\infty} \frac{1}{j^2\pi^2} Z_j^2 \tag{2.16}$$

$$A = \sum_{j=1}^{\infty} \frac{1}{j(j+1)} Z_j^2, \tag{2.17}$$

donde $Z_1, Z_2, ...$ son variables aleatorias i.i.d siguiendo una normal estándar.

En cuanto a las distribuciones exactas de los estadísticos W_n y A_n, a pesar de haber sido largamente estudiadas durante las últimas décadas, los resultados no son tan satisfactorios como los del KS. El estadístico CvM solo ha sido tabulado para $n = 1, ..., 7$, y el AD para un escueto $n = 1$. Sin embargo, se ha visto que la distribución asintótica supone una buena aproximación para valores de n tan pequeños como $n = 8$. El test funcionaría de manera similar al del estadístico KS.

2.4.1. Comentarios finales

Utilizando técnicas de simulación, se ha llegado a la conclusión de que tanto el estadístico CvM como el AD tienen una muy buena potencia en comparación con muchos otros

test. Cuando se quiere contrastar una hipótesis nula simple, su distribución asintótica supone una buena aproximación, incluso para muestras tan pequeñas como $n \geq 10$. Cuando se trabaja con una hipótesis compuesta el procedimiento más adecuado a seguir depende de nuestra hipótesis. Si se desea contrastar la normalidad de la muestra, entonces las distribuciones nulas de AD y CvM han sido tabuladas. En otro caso se recomienda utilizar simulación.

En cuanto al estadístico KS, suele ser menos recomendable por tener menos poder a la hora de detectar la alternativa. Su uso actual en un ámbito especializado debería estar restringido al contraste de una desigualdad estocástica, es decir, $H_0 : F > F_0$ o $H_0 : F < F_0$. A pesar de ello, su popularidad (a menudo unida al PP plot) y sencillez conceptual, justifican su uso cuando no estamos preocupados por realizar test de la máxima potencia posible.

Capítulo 3

Contrastes basados en la estimación de la función de densidad

En este capítulo abordaremos los estadísticos que miden la discrepancia entre una muestra y una distribución hipotética basándose en la función de densidad de esta última. Esto conlleva una restricción obvia: solo es posible aplicar este tipo de tests a distribuciones que admitan una función de densidad, esto es, trabajaremos con variables aleatorias cuya función de distribución $F(X)$ sea absolutamente continua.

El procedimiento es similar al del capítulo anterior en su filosofía. Sea $S_n = x_1, x_2, ..., x_n$ una muestra de n observaciones i.i.d (supondremos que $x_1 < x_2 < ... < x_n$[1]) y sea f su función de densidad. Para contrastar la hipótesis $H_0 : f = f_0$ frente a $H_1 : f \neq f_0$, construiremos estimadores no paramétricos[2] de f a partir de los datos de la muestra y que denotaremos por $\hat{f}_n$. Los estadísticos de contraste funcionarán de nuevo como distancias entre la densidad hipotética y la estimada y buscaremos resultados que nos permitan conocer su distribuciones exactas o asintóticas para poder realizar los constrastes.

Sin bien la idea general es fácil de comprender, especialmente si se ha leido el capítulo de contrastes basados en la función de distribución, adentrarse en la cuestión supone un reto mayor. En primer lugar, existen varios enfoques a la hora de estimar la función de densidad de una muestra. Por limitaciones de tiempo y espacio, hemos decidido centrarnos solo en las estimaciones basadas en funciones kernel. Para una aproximación al problema

[1] Nótese que X es una variable aleatoria continua que admite función de densidad y por ello podemos suponer las desigualdades estrictas. Esto no es cierto en general, y así en el anterior capítulo hemos tenido el cuidado de utilizar desigualdades no estrictas (pues de hecho hemos estudiado cómo aplicar el esdístico KS cuando la función hipotética es discreta)

[2] De forma similar al capítulo anterior, el estimador de f se halla directamente de los datos, sin hacer ninguna suposición previa sobre la distribución que debe seguir. Esto es necesario o del contrario el test carecería de sentido

basada en la aproximación de la f por polinomios ortogonales se puede consultar el manual *Comparing Distributions* de Oliver Thas.

Por otro lado, la literatura enfocada a la estimación de la densidad ha estado históricamente bastante alejada de los contrastes de especificación, además de ser muy reciente y estar en general enfocada a un lector bastante especializado. Así, crear un capítulo estructurado y compacto a nivel de grado ha sido más complicado que en los dos casos anteriores.

3.1. Estimación de la función de densidad

3.1.1. El histograma

El estimador no paramétrico de densidad más antiguo y utilizado es el histograma. Para construirlo se empieza dividiendo la recta real en distintos intervalos que se suelen denominar clases o barras (*bins* en inglés). El histograma es entonces una función escalonada tal que su altura en una clase es la proporción de la muestra contenida en dicha clase dividida por el ancho de la barra (que denotamos por b). La función de densidad estimada toma la siguiente expresión:

$$\hat{f}_H(x;b) = \frac{\text{número de observaciones en la clase que contiene a } x}{nb}. \tag{3.1}$$

A la hora de construir un histograma hay que realizar dos elecciones: la del ancho de barra y la de la posición de las fronteras de las clases. Ambas son significativas por su influencia en la función de densidad estimada. Lo que más nos interesa es el efecto del ancho de barra b, que es un ejemplo de lo que en inglés se suele denominar *smoothing parameter* (que podría ser traducido como «parámetro de suavidad» o «parámetro de fluidez»), y que justamente regula cómo de accidentado se ve el histograma. En general, un ancho de barra pequeño resulta en un histograma más «dentado», mientras que al aumentarlo pasa a ser más «liso». Por analogía, el ancho de barra histograma sirve para ilustrar la importancia del *bandwidth* en los estimadores de tipo kernel que veremos a continuación, ya que precisamente este funciona como un *smoothing parameter* . Otro ejemplo interesante de este tipo de parámetros es el grado de los polinomios elegido para una regresión polinómica, aunque en este caso el *smoothig parameter* sería el inverso del grado.

El histograma tiene varias desventajas que no presentan los estimadores tipo kernel. Uno de ellos es justamente la influencia de la posición de las fronteras de las clases. Otro es que la mayor parte de densidades no son funciones escalonadas, pero los histogramas están limitados a este tipo de funciones. Además, el uso de los datos es menos eficiente que el de los kernel. A pesar de que no vamos a presentar ningún test que esté relacionado

con los histogramas, sí que es una herramienta útil para estudiar la estructura de nuestra muestra y poder tener una idea visual de su densidad real y de las hipótesis nulas que puedan interesarnos. Por ello, junto con su importancia tanto por antigüedad como por popularidad, además de la introducción sencilla que supone a los estimadores de densidad y a los *smoothing parameter*, hemos comenzado el capítulo con esta pequeña sección antes de pasar a los estimadores de verdadero interés en nuestro propósito: los basados en funciones kernel.

3.1.2. Estimación de densidad kernel

En primer lugar veamos qué expresión tiene el estimador de densidad kernel univariante:

$$\hat{f}(x;h) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3.2}$$

Aquí K es una función satisfaciendo $\int K(x)dx = 1$, y a la cual llamamos kernel, y h es un número positivo que cumple una función análoga al ancho de banda del histograma, y por tanto nos referiremos a él por el mismo término. Existe una expresión alternativa para el estimador que resulta de hacer la sustitución: $K_h(u) = h^{-1}K(u/h)$, y así escribiríamos:

$$\hat{f}(x;h) = n^{-1} \sum_{i=1}^{n} K_h(x - x_i) \tag{3.3}$$

Para nuestros propósitos K va a ser una función de densidad unimodal[3] simétrica entorno al punto cero. Aunque se pueden utilizar kernels que no son densidades, esta restricción es interesante para asegurar que nuestro estimador también sea una densidad.

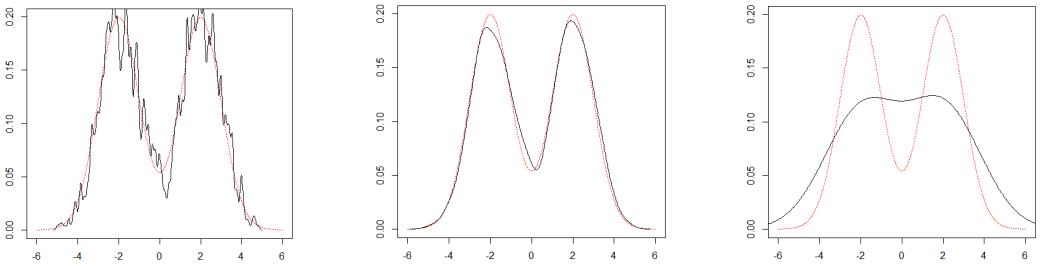
El valor del estimador kernel en un punto x se construye centrando nuestro kernel en cada uno de los datos de la muestra. Desde ahí calculamos qué probabilidades otorga el kernel a las coordenadas de x tomando cada x_i como punto de referencia y hallamos su media. De esta forma, se puede pensar el kernel como una masa de probabilidad de tamaño $1/n$ que toma su máximo en cada dato de la muestra y se extiende de forma simétrica en su entorno. Combinando las contribuciones de cada punto, en las regiones donde hay muchas observaciones el estimador toma valores más altos, como es de esperar.

Aunque pueda parecer sorprendente, la elección de la función de densidad que se emplea como kernel no es demasiado importante (siempre y cuando verifique las restricciones con las que la hemos definido). Sin embargo, sí que tiene una gran importancia el valor h

[3]Dada una función de densidad de una distribución, llamamos modas no estrictas a cada uno de sus máximos locales. En este sentido, una función de densidad unimodal es aquella que tiene un único pico, e.g., la normal

elegido, que como ya hemos adelantado funciona como un *smoothing parameter*. El ancho de banda es en este caso un factor de escala que regula la varianza de nuestro kernel. Cuanto más grande, más «extendida» estará nuestra función K_h, disminuyendo la diferencia en la ponderación de los puntos más cercanos a muchas observaciones de los más alejados. En este caso, el estimador se dice *oversmoothed* y toma una forma muy regular. Por el contrario, un h pequeño acentúa las diferencias de las alturas tomadas en las regiones con más observaciones con las de menos, lo que puede resultar en una excesiva atención a los datos particulares. Esto ocasiona que el estimador tenga una forma más irregular, con muchos máximos y mínimos locales (si h es muy pequeño, cada observación puede corresponderse con un máximo local), y decimos que el estimador está *undersmoothed*.

Ejemplo de la influencia del ancho de banda



(a) $h = 0{,}05$ (b) $h = 0{,}3$ (c) $h = 1{,}5$

Figura 3.1: Las tres imágenes muestran tres estimadores (línea negra) de la densidad de una distribución obtenida de combinar dos normales de medias -2 y 2 y varianza 1 a partir de una muestra simulada de 1000 datos y usando como kernel una función gaussiana. Hemos elegido las medias lo suficientemente alejadas como par que la densidad real (representada por la línea roja punteada) sea bimodal. En el primer caso el estimador está *undersmoothed*, con lo que es muy irregular y presenta numerosos puntos críticos. En el segundo caso tenemos un estimador que podríamos considerar aceptable . En el tercer caso el estimador está *oversmoothed* y, cada vez más regular conforme aumenta h, ya no es capaz de representar el carácter bimodal de la densidad real

Estas cuestiones justifican la necesidad de establecer diferentes medidores de la bondad de ajuste del estimador para la densidad real. Estudiando cómo se comportan estos medidores en función del *bandwidth* o de las propiedades del kernel, se han desarrollado pautas para la elección de la h o, incluso, se ha construido un kernel con propiedades óptimas: el de Epanechnikov. Estas cuestiones las trataremos brevemente al final del capítulo cuando

demos pautas para la implementación práctica del tema. Lo interesante es que estos mismos medidores nacidos desde una perspectiva más descriptiva, han desembocado en los estadísticos de contraste que usaremos en el capítulo, y cuyas propiedades asintóticas se han conseguido desentrañar.

3.1.3. Análisis del estimador

Dado que $\hat{f}_n(x;h)$ es un estimador de $f(x)$ para un $x \in \mathbb{R}$, podemos basarnos en los criterios de error clásicos para ver cómo de precisa es esta estimación. Comenzamos entonces con el error cuadráctico medio[4] (ECM). Para calcularlo hallemos primero la media y la varianza de $\hat{f}_n(x;h)$ a partir de la fórmula 3.3. En primer lugar:

$$E(\hat{f}_n(x;h)) = E(K_h(x-X)) = \int K_h(x-y)f(y)dy \tag{3.4}$$

La expresión de la media motiva la introducción del producto de convolución entre dos funciones, que tiene la siguiente notación:

$$(f*g)(x) = \int f(x-y)g(y)dy \tag{3.5}$$

Una observación que se puede obtener de la expresión de la media es que $\hat{f}_n(x;h)$ es un estimador sesgado de $f(x)$, y su sesgo vale justamente: $(K_h * f)(x) - f(x)$. El producto de convolución puede ser entendido como una transformación de una función en otra más «suave», y así, existe un sesgo que es exactamente la diferencia entre esta versión «suave» de f y la propia f. Esto va a tener importancia más tarde a la hora de elegir estadísticos para el contraste de bondad de ajuste. De momento, continuemos con la expresión de la varianza:

$$Var(\hat{f}_n(x;h) = n^{-1}[(K_h^2 * f)(x) - (K_h * f)^2(x)] \tag{3.6}$$

A partir de ambas expresiones obtenemos la fórmula del ECM:

$$ECM(\hat{f}(x;h)) = n^{-1}[(K_h^2 * f)(x) - (K_h * f)^2(x)] + [(K_h * f)(x) - f(x)]^2 \tag{3.7}$$

El ECM nos aporta información sobre las propiedades puntuales de $\hat{f}_n$ como estimador de la densidad. Sin embargo, al igual que en el capítulo anterior, lo que nos interesa es

[4]Dado un estimador $\hat{\theta}$ de θ definimos su ECM como: $ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$. Una característica muy importante es que puede ser decompuesto en su varianza y su sesgo al cuadrado: $ECM(\hat{\theta}) = Var(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2$

construir medidas del ajuste global. Queremos ver cómo se comporta nuestro estimador en la recta real y no sólo en un punto indeterminado. Es decir, buscamos una distancia funcional, y precisamente emplearemos el cuadrado de la inducida por el producto escalar de L_2 y que da lugar a un medidor muy empleado en estadística paramétrica: el Error Cuadrático Integrado. Este viene dado por la siguiente fórmula:

$$ECI(\hat{f}_n(x;h)) = \int [\hat{f}_n(x;h) - f(x)]^2 dx. \tag{3.8}$$

Este medidor es de utilidad si solo nos preocupa la muestra que tenemos. Para estudiar el comportamiento de nuestro estimador no para un ejemplo concreto, sino para una muestra cualquiera de nuestra densidad f, tendríamos que calcular su esperanza, lo que da lugar al Error Cuadrático Integrado Medio (ECIM),

$$ECIM(\hat{f}_n(x;h)) = E[ECI(\hat{f}_n(x;h))] = E\int (\hat{f}_n(x;h) - f(x)^2 dx \tag{3.9}$$

Haciendo un cambio en el orden de integración obtenemos una fórmula en fuunción del ECM:

$$ECIM(\hat{f}_n(x;h)) = \int E(\hat{f}_n(x;h) - f(x))^2 = \int ECM(\hat{f}_n(x;h)) \tag{3.10}$$

Y aplicando 3.7 obtenemos una expresión muy util del ECIM:

$$ECIM(\hat{f}_n(x;h)) = n^{-1}\int [(K_h^2 * f)(x) - (K_h * f)^2(x)]dx + \\ + \int [(K_h * f)(x) - f(x)]^2 dx \tag{3.11}$$

Podemos notar aquí una consecuencia muy importante del sesgo de $f_n(x;h)$: el ECI no es cero ni tiene a cero conforme n se hace grande; lo cual supondrá un problema más tarde para nuestro estadístico de contraste. Para terminar esta sección, daremos una expresión alternativa de ECIM que proviene de emplear la extensión de $f(x - hz)$ como serie de Taylor y realizar unos cálculos no muy complicados que pueden ser consultados en el manual *Kernel Smoothing* de Wand y Jones. Así, siempre que tenga sentido, tenemos la igualdad siguiente:

$$ECIM(\hat{f}_n(x;h)) = (nh)^{-1}\int K(x)^2 dx + \frac{1}{4}h^4 \left(\int z^2 K(z)dz\right)^2 \int f''(x)^2 dx + o[(nh)^{-1} + h^4].^5 \tag{3.12}$$

A la suma de dos primeros términos de esta expresión se la denomina ECIM asintótico (ECIMA):

$$ECIMA(\hat{f}_n(x;h)) = (nh)^{-1}\int K(x)^2dx + \frac{1}{4}h^4\left(\int z^2K(z)dz\right)^2\int f''(x)^2dx, \quad (3.13)$$

y comportamiento en función de K y h es clave en la búsqueda de estimadores de densidad kernel óptimos.

3.2. Contraste de hipótesis

3.2.1. Estadísticos de contraste

Hasta aquí lo que hemos hecho ha sido presentar la fórmula general para el estimador kernel de densidad y estudiar sus propiedades más importantes. Ahora emplearemos los resultados anteriores para motivar el uso de los dos estadísticos propuestos por Bickel y Rosenblatt y analizarlos para poder entender sus limitaciones.

Como ya se ha comentado en la introducción de este capítulo, el estudio de los estimadores de densidad ha sido históricamente independiente del de los contrastes de especificación, a pesar de su potencial claro como herramienta para este tipo de test. Fueron Bickel y Rosenblatt en su artículo de 1973 quienes por primera vez proponen el uso de estadísticos basados en estimadores de densidad kernel y estudian sus propiedades asintóticas. Lo más interesante es la intuición que tienen de traducir el test de la χ^2 de Pearson (recordamos que estaba basado en el estadístico: $X^2 = \sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i}$) a este nuevo contexto. Así, donde Pearson usa observaciones O_i y valores esperados E_i, Bickel y Rosenblatt los sustituirán por el estimador de densidad $\hat{f}_n(x;h)$ y la densidad hipotética $f_0(x)$. Y como estamos trabajando con variables aleatorias continuas, el sumatorio se traduce en una integral, y en vez de dividir por los valores esperados, añaden una función de peso $w(x)$ a la integral. Es decir, la fórmula general del primer estadístico de Bickel y Rossenblet es:

$$D'_n = \int(\hat{f}_n(x;h) - f_0(x))^2w(x)dx \quad (3.14)$$

En este trabajo tomaremos siempre $w(x) = 1$ y trabajaremos entonces con el estadístico:

[5]Se emplea la notación $a_n = o(b_n)$ cuando $n \to \infty$, si y solo si $\lim_{n\to\infty}|a_n/b_n| = 0$. Se escribe $a_n = O(b_n)$ cuando $n \to \infty$, si y solo si $\lim_{n\to\infty}|a_n/b_n| < \infty$.

$$T'_n = \int (\hat{f}_n(x;h) - f_0(x))^2 dx \tag{3.15}$$

Decidimos esto así por dos motivos. El primero es para evitar extendernos en la cuestión de la elección de $w(x)$ y su efecto en el test, aunque todos los resultados que se presenten más adelante pueden ser generalizados para estadísticos de la forma 3.14. Y el segundo es para poder basarnos en el análisis paramétrico anterior de nuestro estimador de densidad a la hora de estudiar las propiedades de nuestro estadístico T'_n, que bajo la hipótesis nula es exactamente el ICE de nuestro estimador. Así, todos los resultados vistos en la sección anterior se aplican.

Naturalmente, esto implica que T'_n arrastra también la problemática del sesgo de $\hat{f}_n(x;h)$. No debemos olvidar que lo que buscamos es una distancia entre funciones, $d(.,.)$, tal que $H_0 : f = f_0 \Leftrightarrow d(f, f_0) = 0$. En este sentido, $d(\hat{f}_n, f_0)$ es un estimador de $d(f, f_0)$, y por tanto nos interesa que tenga buenas propiedades, como que sea insesgado o consistente. La fórmula 3.11 nos da una expresión de la esperanza del estadístico bajo la hipótesis nula:

$$E(T'_n) = n^{-1} \int [(K_h^2 * f_0)(x) - (K_h * f_0)^2(x)]dx + \int [(K_h * f_0)(x) - f_0](x)]^2 dx \tag{3.16}$$

Vemos claramente que la esperanza de T'_n es estrictamente mayor que 0, y que, de hecho, no tiene a 0 cuando n se hace más grande. Es decir, T'_n no es un estimador insesgado ni asintóticamente insesgado de $d(f, f_0)$ bajo la hipótesis nula, lo cual es consecuencia del sesgo de $\hat{f}_n$. Existen dos formas de solventar este defecto. La primera y más obvia es sustituir el estadístico T'_n por uno que tenga mejores propiedades; en particular, que sea asintóticamente insesgado. Como el problema siempre es la diferencia entre el producto de convolución de nuestra función hipotética y la propia función, la solución más inmediata consiste en sustituir f_0 por $K_h * f_0$ en la definición de nuestra T'_n, lo que da lugar al nuevo estadístico T_n con fórmula:

$$T_n = \int [\hat{f}_n(x;h) - (K_h * f_0(x))]^2 dx, \tag{3.17}$$

que también podemos escribir como:

$$T_n = \int [\hat{f}_n(x;h) - E_{H_0}(\hat{f}_n(x;h))]^2 dx, \tag{3.18}$$

donde $E_{H_0}(\hat{f}_n(x;h))$ indica la esperanza de nuestro estimador de densidad bajo la hipótesis nula. Este estadístico es propuesto de una forma más general por Bickel y Rosenblatt

al incluir también una función de peso, pero como ya hemos indicado preferiremos tomar siempre como peso la unidad y así obviar el problema de su elección.

La segunda solución consiste en hacer tender nuestro ancho de banda a 0 conforme aumenta el tamaño de la muestra. Efectivamente, si en la expresión 3.12 sustituimos la f por nuestra f_0 tendremos la otra expresión de la media de T'_n bajo la hipótesis nula, siempre y cuando se cumplan ciertas condiciones de regularidad.

$$E(T'_n) = (nh)^{-1} \int K(x)^2 dx + \frac{1}{4} h^4 \left(\int z^2 K(z) dz \right)^2 \int f''(x)^2 dx + o[(nh)^{-1} + h^4]. \tag{3.19}$$

Lo interesante de esta expresión es que depende de nuestra h de una forma muy simple y nos permite ver que si conseguimos hacer tender tanto $(nh)^{-1}$ como h a 0, tendremos que T'_n sí será asintóticamente insesgado. Además, este expresión nos ayudará a entender el motivo de algunas de las restricciones que precisaremos para poder asegurar la convergencia de $\hat{T}_n$ que estudiaremos ahora. Lo que podemos adelantar es que necesitaremos tomar una sucesión de anchos de banda h_n (por comodidad seguiremos empleando h para denotar $h = h_n$) tal que $\lim_{n\to\infty} h = 0$ y $\lim_{n\to\infty} (nh)^{-1} = 0$.

3.2.2. Comportamiento asintótico

Hasta ahora hemos presentado la fórmula general de un estimador de densidad kernel, hemos empleado herrramientas del análisis paramétrico para llegar a la expresión de los dos estadísticos que estudiaremos en este capítulo y hemos obtenido ciertas restricciones que nos aseguran un mejor comportamiento de al menos uno de ellos. Hemos intentado que todo ello fuese de forma fluida y, al mismo tiempo, todo lo riguroso que hemos sabido.

Pero este enfoque tiene que ser abandonado en este punto. Queda por ver cuáles son las distribuciones asintóticas de nuestros estadísticos de contraste bajo la nula, de forma que nos sirvan para hacer inferencia. El problema es que las demostraciones toman una complejidad bastante elevada, tanto por su tamaño como por la base teórica que requieren, y por ello serán pasadas por alto. Solo daremos entonces las condiciones que necesitaremos para asegurar sus convergencias (muchas de ellas están estrechamente relacionadas con la fórmula 3.19 como se podrá ver), y sus propias distribuciones límite.

Añadamos entonces las siguientes restricciones a las ya dadas:

1. Nuestro ancho de banda es una sucesión $h = h_n$ tal que $\lim_{n\to\infty} h = 0$ y $n^{-1} = o(h)$.

2. K cumple o que (a) es nula fuera de un intervalo $[-A, A]$ y absolutamente continua

en $[-A, A]$, o bien que (b) es absolutamente continua en la recta real y $\int |K''|^k < \infty$ para $k = 1, 2$.

3. La densidad f es continua, positiva y acotada. Y $f^{1/2}$ es absolutamente continua y su derivada $\frac{1}{2}f'/f^{1/2}$ está acotada. Y además:

$$\int_{[|z|^{3/2}\geq 3]} |z|^{3/2}[log\ log|z|]^{1/2}[|K'(z)] + |K(z)|]dz < \infty \tag{3.20}$$

4. La segunda derivada f'' de f existe y está acotada.

5. Existe $\int z^2 K(z)dz$

Una vez expuestas estas restricciones podemos comenzar por el estadístico T_n. Definimos $\mu(K) = \frac{1}{nh}\int K^2(z)dz$ y $\sigma(K)^2 = 2[\int(\int K(u+v)K(v)dv)^2du]\int f_0^2(x)dx$. Se cumple entonces el siguiente teorema:

Teorema 3.1. *Supngamos que se cumplen las restricciones 1-3, entonces, bajo la hipótesis nula y cuando $n \to \infty$:*

$$n\sqrt{h}\frac{T_n - \mu(K)}{\sigma(K)} \xrightarrow{d} N(0,1) \tag{3.21}$$

Para presentar este resultado hemos hecho uso de solo las tres primeras restricciones que hemos presentado, lo cual tiene sentido ya que T_n es un estadístico con mejores propiedades que T_n'. Nótese además en 3.19, que necesitamos las restricciones 4 y 5 para que $\hat{T}_n$ tenga esperanza. Sin embargo, el estadístico T_n' es probablemente de mayor interés que T_n, y gracias a las restricciones 4 y 5, podemos asegurar también su convergencia a una normal de la que indicaremos la media y la varianza en el siguiente teorema.

Teorema 3.2. *Supongamos que se cumplen 1-5. Entonces, bajo la hipótesis nula y cuando $n \to \infty$:*

$$n\sqrt{h}\frac{T_n' - \mu(k)}{\sigma(k)} \xrightarrow{d} N(0,1) \tag{3.22}$$

Es decir, siempre que se cumplan las restricciones 1-5, los dos estadísticos tienen el mismo comportamiento asintótico. El contraste de la hipótesis simple $H_0 : f = f_0$, con un nivel de significación α lo haríamos de la siguiente forma: hallamos el valor de T_n a partir de la muestra, calculamos $d(\alpha) = \mu(K) + (n\sqrt{h})^{-1}\sigma(K)\phi(1-\alpha)$ y, si $T_n > d(\alpha)$, rechazamos la hipótesis nula. En caso contrario la aceptamos.

3.2.3. Hipótesis compuesta

Si el uso del estimador de densidad kernel para el contraste de especificación de una hipótesis simple es de mediados de la década de los ochenta, su estudio para el contraste para una hipótesis compuesta $H_0 : f \in \{f_\theta : \theta \in \Theta\}$ es aún más reciente. En efecto, Bickel y Rossenblet no estudian el comportamiento de los estadísticos T'_n y T_n cuando sustituimos nuestra densidad f_0 por una $f_{\hat{\theta}}$ con $\hat{\theta}$ un estimador de θ obtenido a partir de la muestra. Ya hemos visto en el primer capítulo (a raíz de la sonada controversia entre Pearson y Fisher) que las distribuciones límites no tienen por qué ser en general las mismas, pues la inclusión del parámetro estimado dependiente de la muestra puede tener un efecto sobre su comportamiento asintótico.

De nuevo, vamos a evitar entrar en demostraciones demasiado laboriosas e iremos directamente a los resultados que nos interesan. Solo vamos a presentar un teorema sacado de Fan (1994), quien propone emplear el estadístico $\hat{T}_n$ que resulta de sustituir f_0 por $f_{\hat{\theta}}$ en la fórmula de T_n, y donde $\hat{\theta}$ lo hallamos empleando el método de máxima verosimilitud. Es decir:

$$\hat{T}_n = \int (\hat{f}_n(x;h) - (K_h * f_{\hat{\theta}})(x))^2 dx \tag{3.23}$$

Bajo ciertas hipótesis de regularidad que no introduciremos (se puede consultar en Fan (1994)) y bajo la hipótesis nula, tenemos que, cuando $n \to \infty$:

$$n\sqrt{h}\frac{\hat{T}_n - \mu(k)}{\sigma(K, \hat{f}_n))} \xrightarrow{d} N(0,1), \tag{3.24}$$

donde ahora

$$\sigma(K, \hat{f}_n)^2 = 2\left[\int \left(\int K(u+v)K(v)dv\right)^2 du\right] \int \hat{f}_n(x;h)^2(x)dx$$

El test resultante rechaza la hipótesis nula $H_0 : f \in f_\theta : \theta \in \Theta$ cuando $\hat{T}_n > d(\alpha)$ con $d(\alpha) = \mu(K) + (n\sqrt{h})^{-1}\sigma(K)\phi^{-1}(1-\alpha)$.

3.3. Comentarios finales

Para finalizar este capítulo vamos a comentar ciertas cuestiones importantes que hay que tener cuenta a la hora de implementar los test en la práctica.

En primer lugar, apenas se ha hablado de la influencia de la función kernel elegida para el contraste, aunque sí que hemos adelantado que el comportamiento del estimador

es bastante similar independientemente de la K escogida. Se ha demostrado que el kernel de Epanechnikov, que denotamos por K^*, es el que minimiza el ECIMA y suele ser usado como referencia para estudiar el comportamiento de otros kernels propuestos. Decimos que el kernel K tiene una eficiencia el 90 % si el estimador de densidad óptimo (el que emplea K^* como kernel) puede conseguir el mismo ECIMA utilizando el 95 % de los datos que usando K. Un estudio de la eficiencia de distintas funciones muestra que los kernels clásicos, entre los que se encuentra la normal o la uniforme, tienen un rendimiento superior al 90 %. Por ello la elección suele atender a criterios computacionales, y kernels como el uniforme (que es constante a trozos) o incluso el de Epanechnikov (cuya derivada es discontinua) suelen ser rechazados por otros más regulares.

En segundo lugar, queda el problema de la elección del ancho de banda. Estudios por simulación de los test basados en estimadores de densidad kernel han mostrado una gran dependencia del *smoothing parameter*, por lo que es una cuestión clave. Sin embargo, también es una cuestión de muchísima complejidad. Se han propuesto numerosos criterios para definir una expresión de h que resulte en buenas propiedades para nuestro estimador (por suepuesto, todas estas expresiones cumplen que h tiende a cero conforme el tamaño de la muestra se hace grande). Sin embargo, su éxito en la reducción del ECIMA depende en gran medida de la función de densidad que siga la muestra, con lo que es difícil establecer un criterio general. Además, como estas cuestiones nos interesan desde el punto de vista de los contrastes de especificación, existe la dificultad añadida del hecho de que muchos de estos criterios hagan depender a h no solo de n, sino de los propios valores de la muestra. Esto crearía una nueva dependencia de nuestros estadísticos sobre la muestra, y no tenemos la seguridad de la validez de los teoremas presentados sobre su comportamiento asintótico. De esta forma, la elección óptima del parámetro es una cuestión que continúa hoy en día sin tener una respuesta clara.

Capítulo 4

Ilustración en bases de datos simulados

Para finalizar el trabajo, vamos a llevar a la práctica los test de contraste vistos en los tres capítulos anteriores. Usaremos para ello bases de datos simulados, lo que nos permitirá poder interpretar el resultado de los test sabiendo en todo caso si nuestra hipótesis nula es cierta o no. Es decir, sabremos si el test es acertado o está cometiendo algún tipo de error.

Es necesario aclarar que en ningún caso pretendemos hacer un estudio pormenorizado y sistemático de los estadísticos de contraste. No buscamos concluir qué estadístico es el más adecuado, qué número de intervalos es el óptimo para el test de Pearson o cuál es el ancho de banda idóneo para nuestro estimador de densidad. Para ver estudios verdaderamente rigurosos a partir de simulación nos remitimos a la bibliografía. En este trabajo hemos comentado muchas de las conclusiones de estos estudios (así como cuestiones que están lejos de estar concluidas), y no pretendemos verdaderamente añadir ni modificar nada.

Nuestro objetivo es otro, y, además, es doble. En primer lugar, y especialmente, buscamos ejemplificar todos los métodos de los que hemos hablado, de forma que su comprensión teórica se complete con un caso aplicado a una muestra concreta. De poco sirve comprender la teoría que lleva a la identificación del proceso empírico en un proceso Gaussiano, si no sabemos cómo realizar un contraste empleando el estadístico de Kolmogorov-Smirnov. Este trabajo trata en último lugar sobre contrastes, y aunque nos hayamos centrado en los aspectos más teóricos, pues nos interesaba sobretodo hacer una revisión de las familias de estadísticos más importantes y de las herramientas matemáticas en las que se basan, no podemos olvidar el problema no tan obvio de cómo aplicar la teoría en un caso práctico.

Por otro lado, este capítulo nos servirá para seguir introduciendo conceptos y técnicas de la estadística computacional tan fundamentales como el método Monte Carlo o el

Bootstrap paramétrico. Nuestro acercamiento a estos métodos será sobretodo intuitivo y a partir de ejemplos. Los iremos motivando a través de las diferentes cuestiones que nos irán surgiendo sobre las propiedades de nuestros estadísticos al trabajar con nuestros datos simulados. Un enfoque más riguroso ocuparía demasiado espacio y haría que nos desviáramos excesivamente del tema que nos ocupa.

De ahora en adelante daré una descripción de los métodos implementados en R para el contraste e incluiré los resultados obtenidos seguidos de las conclusiones que podemos sacar. El código lo presentaremos al final del trabajo en un apéndice.

4.1. Hipótesis simple

Vamos a simular en R un muestra S_n de una normal estándar tamaño n que de momento no fijaremos pues nos interesará variarla en ocasiones. Nuestra hipótesis nula será correcta, $H_0 : S_n \in N(0,1)$ frente a $H_1 : S_n \notin N(0,1)$, con lo que podremos estudiar el error de tipo I de nuestros test.

4.1.1. Contraste usando los estadísticos de divergencia

Comenzamos por los test del primer capítulo. La primera cuestión es cómo discretizar la normal. Como estamos en el caso de la hipótesis simple, podemos aplicar el criterio de equiprobabilidad. Además vamos a tomar siempre $n \geq 50$, con lo que podremos tomar $k = 10$ intervalos equiprobables y de forma que las observaciones esperadas en cada intervalo sean siempre mayores o iguales que 5, como recomendábamos para que la χ^2 sea una buena aproximación de la distribución exacta de nuestros estadísticos. Vamos a trabajar con tres estadísticos: el X^2 de Pearson, el G^2 del test de razón de verosimilitudes y el $2nI^\lambda$ con $\lambda = 2/3$, que es el sugerido como óptimo por Read y Cressie de entre todos los estadísticos de la familia de divergencia y que denominamos PD. Como tienen expresiones analíticas podemos calcularlos con facilidad (existen funciones de paquetes de R pero solo las usaremos para comparar). Usaremos siempre una *seed* determinada de forma que nuestros resultados puedan ser más tarde corroborados. Fijaremos siempre $set.seed(3000)$ y tomamos $n = 50$. Para hallar el p-valor utilizamos la distribución asintótica χ^2_9: para el estadístico X^2 lo calcularíamos como $1 - F_{\chi^2_9}(X^2)$ siendo $F_{\chi^2_9}(X^2)$ la función de distribución de la χ^2_9. Para G^2 y PD se haría de forma similar. Los resultados los recogemos en la tabla 4.1.

Tomando cualquier nivel de significación clásico y aplicando cualquiera de los test aceptaríamos la hipótesis nula. Nótese que los tres valores son muy similares (como hemos probado). Este sería el procedimiento que deberíamos para realizar el contraste. Por

	Estadístico		
	X^2	G^2	PD
Valor	$12,000$	$11,860$	11,751
p-valor	0,213	0,221	0,228

Cuadro 4.1: Tabla con el valor de los estadísticos

supuesto, en R ya existen funciones incluidas en paquetes que realizan estos test, y que recomendamos utilizar. Se puede comprobar sin embargo que los resultados obtenidos son exactamente los mismos.

Podemos sin embargo intentar ir un paso más allá. Este resultado concreto parecería indicar que los tres test funcionan correctamente a la hora de reconocer la función hipotética, pero una sola muestra no parece suficiente para poder sacar conclusiones sobre el error de tipo I. Una idea interesante sería repetir este proceso un número muy grande de veces y comparar cuántas veces la hipótesis nula es rechazada por cada tipo de test respecto del total. Necesariamente tenemos que olvidarnos de fijar la *seed*, pues queremos muestras que puedan ser diferentes sacadas de la normal estándar. Lo haremos para 100000 repeticiones y haciendo variar la n.

Surge aquí un problema con el estadístico de razón de verosimilitudes: no tiene valor cuando algún intervalo no tiene valores observados. Cuando n es grande este inconveniente desaparece pues siempre habrá alguna observación en alguno de los 10 intervalos, pero cuando $n = 50$ este problema se da de forma ocasional y si usamos varios miles de muestras va a haber sin duda casos en los que no esté definido. Lo ideal sería poder elegir intervalos equiprobables no vacíos, pero hacerlos depender de la muestra es arriesgado como ya hemos comentado. Para no entrar en mayores complicaciones y como este problema solo existe si n es bajo, vamos a calcular de igual forma cuántas veces el test da positivo de todas las veces en las que esté definido, e indicaremos en una nota a pié de página cuántas veces no se ha podido definir. Lo haremos para $n \in \{50, 100, 500, 1000\}$. Los resultados están recogidos en la tabla 4.2.

Como era de esperar, el número de veces que los estadísticos han aceptado el test es similar, y parece tender a un 95 % del total. Esto es se explica porque el nivel de significación que hemos usado es $\alpha = 0{,}05$. Los test dan positivo siempre que el estadístico de contraste es inferior al cuantil de orden 95 de la χ^2_9, con lo que estamos comprobando que este valor es muy cercano al cuantil de orden 95 de las distribuciones exactas de nuestros estadísticos según cada n. Esto es lógico, pues por la teoría asintótica vista sabemos que

	Estadístico		
	X^2	G^2	PD
n=50	0,953	0,960[1]	0,953
n=100	0,951	0,944[2]	0,952
n=500	0,950	0,949	0,950
n=1000	0,952	0,951	0,952

Cuadro 4.2: Proporción de veces que el test es positivo bajo la hipótesis nula

estos estadísticos convergen en distribución a una χ^2_9.

Lo que acabamos de hacer no es sino utilizar el método de Monte Carlo para hallar el error de tipo I empírico de nuestros test. Basándonos en él, podemos concluir que a partir de $n = 50$ el error de nuestros test es muy cercano al nivel de signficación pedido, y a partir de $n = 100$ casi exactamente el mismo. Aún más, podemos generalizar este razonamiento sobre el cuantil y directamente querer contrastar si la χ^2 es una buena aproximación de la distribución exacta de nuestros estadísticos para un determinado n. Gracias al método de MonteCarlo contamos con una muestra de tamaño 100000 de cada uno de estos estadísticos bajo la hipótesis nula. Utilizando por ejemplo el test de Kolmogorov-Smirnov, podemos comparar cuál está más cercano a la distribución hipotética. Los p-valores resultantes están recogidos en la tabla 4.3.

	Estadístico		
	X^2	G^2	PD
n=50	$2{,}2e-16$	$2{,}2e-16$	$2{,}2e-16$
n=100	$2{,}2e-16$	$2{,}2e-16$	$1{,}6e-7$
n=500	0,029	0,650	0,549
n=1000	0,343	0,604	0,712

Cuadro 4.3: p-valores resultantes de testear con el test de Kolmogorov-Smirnov (comando *ks.test()*)la adecuación de los estadísticos a la χ^2

Podríamos de aquí concluir que en general el estadístico PD sí parece ser el que mejores propiedades tiene, con un error de tipo I empírico muy parecido al pedido y teniendo una distribución exacta más parecida en general a la asintótica que la de los otros dos estadís-

[2]Exactamente 5062 de entre las 100000 muestras han dejado algún intervalo vacío y por tanto el estadístico G^2 no está definido. Solo ha ocurrido este incidente para este valor de n

[2]Veinte valores indeterminados

ticos. El test de razón de verosimilitudes también parece tener un buen comportamiento, sin embargo el problema de en ocasiones no estar definido justifica que fuera rechazado por Read y Cressie. De cualquier manera, como indicábamos en la introducción, no buscamos sacar conclusiones rigurosas, sino que lo que nos interesa verdaderamente son los métodos que estamos implementando durante el proceso seguido.

4.1.2. Contraste con los estadísticos basados en el proceso empírico

Ahora procederemos a hacer el contrate usando los estadísticos del segundo capítulo. Fijamos $n = 100$ y $set.seed = 3000$ y simulamos la muestra. Primero de todo nos puede interesar comparar la distribución empírica con la hipotética, con lo que representamos ambas gráficamente 4.1.

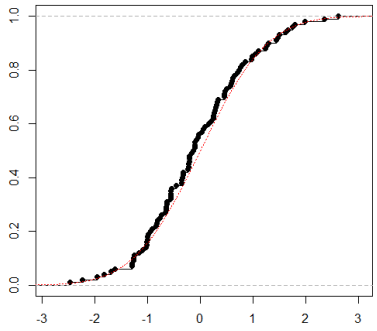


Figura 4.1: La distribución empírica se asemeja mucho a la de la normal estándar.

Trabajaremos con los tres test introducidos el segundo capítulo: el de Kolmogorov-Smirnov, el de Cramér-von Mises y el de Anderson-Darling. Para ello utilizaremos nosotros nuestro propio código basado en las fórmulas dadas en el capítulo. En R existe el comando *ks.test*, que permite efectuar el test de KS de una forma rápida; y también nos será muy útil el paquete *goftest*, ya que cuenta con funciones para realizar tanto el test de CvM como el de AD. No emplearemos estos test para realizar directamente el contraste, nuestro objetivo es llevar la teoría a la práctica, y ello implica intentar computar las expresiones que hemos introducido. Consideramos que empleando funciones cuyo código desconocemos restaríamos mucho valor a este capítulo, con lo que lo evitaremos dentro de lo posible. Sin embargo, sí que emplearemos estas funciones ya hechas para contrastar nuestros resultados, y confirmar así tanto que nuestro código está bien construido, como que la base teórica de los test que continuamente empleamos hoy en día se corresponde con la expuesta en este trabajo.

Procedemos entonces a hallar el valor de los estadísticos para la *seed* y el n antes fijados y que exponemos en la tabla. Para calcular el valor del estadístico de KS lo expresamos como: $D_n = \underset{1\leq i\leq n}{max}\{D_n^+, D_n^-\}$ y echamos mano de las expresiones 2.3 y 2.4. Para realizar el contraste utilizamos la fórmula analítica en serie de potencias de su distribución límite nula 2.7, aproximándola, por ejemplo, mediante sus mil primeros sumandos. El p-valor lo calculamos como: $1 - F_D(D)$.

Trabajar con los estadísticos de CvM y AD supone un reto algo mayor ya que no contamos con su distribución exacta para un n tan alto, y solo conocemos su distribución asintótica como una suma infinita ponderada de normales estándares al cuadrado. Naturalmente, no conocemos la distribución exacta de dicha suma de variables aleatorias, con lo que en un principio no podríamos realizar el contraste. En realidad, esto es algo muy común en numerosos ámbitos de la estadística. Si solo pudiéramos trabajar con estadísticos que tengan como distribución exacta o asintótica una distribución clásica veríamos muy restringidas nuestras posibilidades. Por suerte, contamos con la herramientas de la simulación para generar varias muestras que provengan de esta suma de normales al cuadrado, y emplear el método de Monte Carlo para hallar el cuantil empírico de orden 95. Simulando 10000 veces esta suma hasta su centésimo término, hemos obtenido unos valores de: 0,4671 y 2,5097 para los cuantiles de las distribuciones asintóticas nulas de CvM y AD respectivamente. Además, hemos podido construir sus funciones de distribución empíricas, que nos permitirán hallar los p-valores. Usando estos valores y calculando los estadísticos usando las expresiones 2.12 y 2.13 podemos llevar a cabo el test.

	Estadísticos		
	D_{100}[3]	W_{100}	A_{100}
Valor	0,072	0,099	0,072
p-valor	0,671	0,581	0,752

Cuadro 4.4: Tabla con nuestros resultados para el valor de los estadísticos y los p-valores

Observando los resultados de 4.4 deducimos que realizando cualquiera de los tres test habríamos aceptado la hipótesis nula. Comprobando los resultados con las funciones de R (tabla4.5) vemos que son prácticamente idénticos y nos aseguramos de que nuestro código funciona correctamente. Como en el capítulo anterior, nos interesamos también por ver el comportamiento del error de tipo I para muchas muestras simuladas y medimos con el

[3]Expresamos el estadístico de Kolmogorov-Smirnov sin multiplicar por $\sqrt{n}$ para poder compararlo mejor con el resultado del comando ks.test()

	Test		
	ks.test	cvm.test	ad.test
Valor	0,072	0,100	0,488
p-valor	0,672	0,584	0,758

Cuadro 4.5: Tabla con los resultados de los test de R

test de Kolmogorov-Smirnov si la distribución de los D_n se acerca a F_D, y si las de A_n y W_n se acercan a las empíricas de la suma que hemos computado anteriormente. Para ello, retiraremos la semilla y simularemos mil muestras para cada $n = 50, 100, 500, 1000$. Los resultados están en las tablas 4.6 y 4.7.

	Estadístico		
	D_n	W_n	A_n
n=50	0,958	0,948	0,950
n=100	0,951	0,952	0,960
n=500	0,967	0,942	0,956
n=1000	0,953	0,955	0,957

Cuadro 4.6: Proporción de veces que el test es positivo bajo la hipótesis nula

	Estadístico		
	D_n	W_n	A_n
n=50	0,0004	0,074	0,050
n=100	0,192	0,339	0,374
n=500	0,799	0,576	0,374
n=1000	0,981	0,269	0,591

Cuadro 4.7: p-valores resultantes de testear con el test de Kolmogorov-Smirnov la adecuación de los estadísticos a su distribución asintótica

Los resultados son satisfactorios. El error empírico es cercano al precisado, aunque quizás nos gustaría una convergencia más clara al valor 0,950 conforme aumente n con en el caso de la sección anterior. El hecho de que los resultados del test de Kolmogorov-Smirnov para medir la adecuación de los estadísticos a sus distribuciones asintóticas sea tan concluyentes a favor de esta premisa, nos invita a plantearnos que quizás el problema

pueda estar más bien en la estimación del cuantil empírico que hemos tenido que realizar para W y A. Un análisis mucho más detallado y con mucha más potencia computacional sería necesario para poder sacar conclusiones.

4.1.3. Contrastes con los estadísticos basados el la estimación de la función de densidad

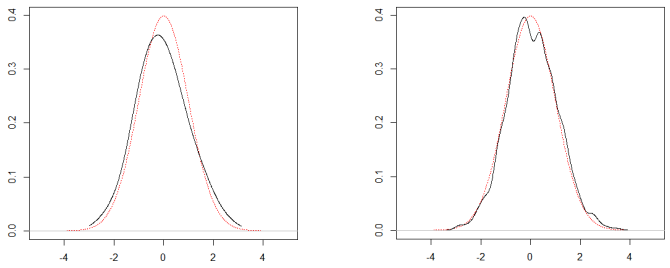
Finalmente, emplearemos los estadísticos del tercer capítulo para realizar el contraste. Podemos comenzar con $n = 100$, fijando $set.seed(3000)$ y simulando nuestra muestra. Una primera aproximación a su densidad es el histograma que construimos con $b = 0{,}5$ y representando las frecuencias relativas. Este ya nos parece indicar que la densidad de la muestra es unimodal y simétrica en torno al punto 0, apuntando a una posible distribución normal.

Directamente podemos pasar ya a hallar un estimador de densidad kernel. Como no tenemos pautas para la elección del h, vamos a coger $h = 0{,}5$, y como la función que tomemos no tiene demasiada importancia, emplearemos el kernel gaussiano que viene implícito en la función *density* de R y que es mejor que otras opciones por ser más regular. Hemos escogido este valor de h porque en general hemos visto que da buenos resultados, aunque podríamos haber elegido cualquier otro. Lo que no podemos es elegir h en función de cómo de bien el estimador se adapta a la densidad real, ya que no podemos hacer depender el ancho de banda de la muestra o nuestra teoría asintótica no tendría validez.

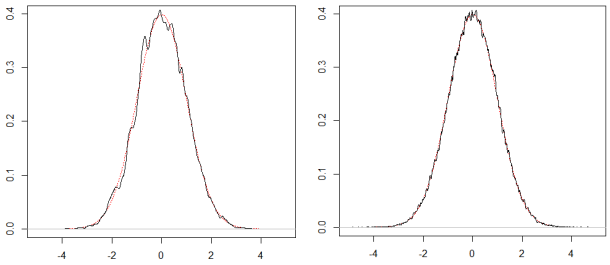
En cambio, sí que tendremos que hacer depender la h del tamaño de la muestra, de forma que tienda a 0 cuando esta se haga grande y se cumpla que $n^{-1} = o(h)$. La forma más sencilla de hacer cumplir esto sería tomando $h = a/\sqrt{n}$. Como para $n = 100$ elegimos $h_{100} = 5$, tomamos $a = 5$. Una vez que hemos hecho esta elección sí podemos tener la curiosidad de querer comprobar visualmente cómo se comportan estos estimadores respecto de la densidad real para distintos n, lo cuál es posible en la figura 4.2.

Ahora procederíamos como en los dos casos anteriores. Pero nos encontramos con dos dificultades. La primera es que nuestros estadísticos no tienen expresiones analíticas en general, con lo que tenemos que emplear los métodos de integración numéricos de los que dispone R. La segunda es que no hemos conseguido encontrar paquetes con el test de Bickel y Rosenblatt. Aunque hasta ahora no hemos empleado los paquetes directamente para realizar los contrastes, sí los hemos usado para comprobar que nuestros test están bien definidos y ver así que las funciones que tanto se emplean en R están efectivamente asentadas en toda la teoría descrita en este trabajo.

A pesar de todo ello, vamos a intentar aplicar los métodos de igual manera, siendo



(a) $n = 100$ (b) $n = 1000$



(c) $n = 10000$ (d) $n = 100000$

Figura 4.2: Las cuatro gráficas representan el estimador kernel con las condiciones descritas arriba y la función de densidad real. Como vemos conforme aumenta el n mejor ajustada parece estar la estimación, aunque también es más irregular. Hay que tener en cuenta que al estar trabajando con una normal (función unimodal y simétrica), el método funciona mucho mejor que para funciones más irregulares.

conscientes de nuestras limitaciones. No vamos a comprobar tampoco que se cumplen las 5 hipótesis que numerábamos en el capítulo 3. Estamos empleando la normal estándar tanto como kernel como función de densidad hipotética (que sabemos que es la real), y su regularidad está asegurada. Así, pasamos a calcular el valor de los estadísticos según nuestra muestra y empleamos su distribución normal asintótica para realizar el contraste. El resultado se recoge en la tabla 4.8, y podemos ver como el test da positivo, con p-valores muy similares.

	Estadístico	
	T'	T
Valor	0,00474	0,00214
p-valor	0,575	0,550

Cuadro 4.8: Tabla con el valor de los estadísticos

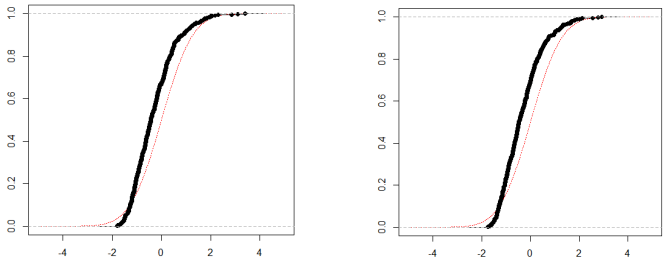
Para analizar el error de tipo I de nuestros estadísticos usamos Monte Carlo: simulamos muchas muestras y hallamos el error de tipo empírico. Nótese que este método es probablemente el más costoso computacionalmente de todos, no tanto por el número de veces que tenemos que realizar aproximaciones a integrales, como por la necesidad de hallar el estimador kernel de cada muestra. Como nos interesa que sea una función y no un vector, hemos empleado el comando *kdensity* del paquete *kdensity*, pero esto conlleva un coste alto. Vamos a repetir el método 500 veces, para lo que necesitamos ya bastante potencia dependiendo de qué n tomemos. Lo haremos de nuevo para $n = 50, 100, 500, 1000$. Los resultados los recogemos en la tabla 4.9.

	Estadístico	
	T'	T
n=50	0,926	0,994
n=100	0,95	0,994
n=500	0,974	0,98
n=1000	0,97	0,972

Cuadro 4.9: Errores de tipo I empíricos

Los resultados son ligeramente insatisfactorios, pues aunque el error empírico es bastante cercano al teórico, no parece que conforme aumente el tamaño de la muestra este vaya acercándose a 0,95, como es esperable teniendo en cuenta la convergencia en distribución de los estadísticos. También hemos realizado el test de KS para testear normalidad y siempre hemos obtenido p-valores muy bajos, con lo que rechazaríamos en todo caso que los estadísticos pudieran ser aproximados correctamente por su distribución asintótica. Por tener una idea visual de cómo de cerca se encuentra la distribución exacta de la hipotética, hemos computado la distribución empírica y la hemos representado gráficamente junto con la nula para $n = 1000$.

Este comportamiento ligeramente alejado del teórico quizás se deba a las limitaciones con las que nos encontramos para poder realizar los test, como el tamaño modesto de la muestra o el número de repeticiones. Seguramente también tenga que ver la influencia del ancho de banda, que ya habíamos adelantado que juega un papel determinante en el comportamiento de los estadísticos, pero para el que no contamos con criterios para optimizarlo. Teniendo en cuenta estas cuestiones, los resultados que hemos obtenido pueden darse por satisfactorios dentro de las aspiraciones de nuestro trabajo.



(a) $n = 100$ (b) $n = 1000$

Figura 4.3: Ambas distribuciones parecen bastante cercanas a la normal. Sin embargo, teniendo en cuenta que estamos ante un valor de n grande, el test de kolmogorv-smirnov es muy sensible a desviaciones de lo esperado, y por tanto los p-valores son muy pequeños. Por otro lado, el test sí que ha aceptado ambos estadísticos siguen la misma distribución exacta con un p-valor muy alto (p=0.82)

4.2. Hipótesis compuesta

Vamos a ejemplificar ahora el contraste suponiendo que tenemos una hipótesis compuesta. De nuevo generaremos datos de una normal estándar (elegimos de nuevo la normal porque ello nos permitirá emplear el test de Lilliefors del segundo capítulo), y consideraremos la hipótesis nula correcta, en este caso: $S_n \in \{N(\mu, \sigma) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$. De esta forma podremos estimar de nuevo el error empírico de tipo I de nuestros test.

Todos los test de nuestros capítulo precisan la estimación de los parámetros de la normal a partir de la muestra para poder llevarse a cabo. Vamos a emplear el método de razón de verosimilitudes para realizar esta estimación, ya que es el más universalmente extendido y el único para el que tenemos la garantía de todos los resultados asintóticos de nuestros estadísticos para la hipótesis compuesta. Además, contamos con la función *mle* del paquete *stats4* para poder realizar la estimación. Recordamos sin embargo que este no es el único estimador que existe, ni el único con buenas propiedades (recuérdese la familia de estimadores BAN a la que pertenecían los estimadores asociados a los estadísticos de divergencia), pero varios de los teoremas que hemos presentado aseguran la convergencia solo para el estimador máximo verosímil.

De esta forma, fijemos la semilla $set.seed(3000)$ con $n = 100$ y generamos la muestra. Utilizamos la función *mle* y obtenemos $\hat{\mu} = -0{,}033$ y $\hat{\sigma} = 0{,}924$, efectivamente los parámetros estimados están muy cerca de los reales. Ahora, en vez de proceder como en la sección anterior y separar los contraste por capítulos, como muchas de las cuestiones relativas a los contrastes ya han sido tratadas vamos a hacer una sección más compacta para poder

ya centrarnos en la comparación entre los test, aunque comentando igualmente las dudas que puedan surgir sobre los métodos aplicados al caso compuesto.

Comenzando por los estadísticos de divergencia, la principal diferencia en el método respecto a la anterior sección es que ya no podemos tomar intervalos equiprobables (o por lo menos no sin que esta elección dependa de la muestra). El único criterio que podíamos seguir entonces era situar las fronteras de los intervalos de forma aleatoria. Como necesitamos que haya por lo menos un dato por intervalo, usaremos la distribución uniforme para obtener nueve puntos entre el máximo y el mínimo de nuestra muestra. Obviamente esto hace depender nuestros intervalos de los datos, pero utilizar un método aleatorio cualquiera probablemente inhabilitaría totalmente nuestro test ya que clasificaría todos los datos en un mismo intervalo. Tomaremos de nuevo $k = 10$ y, salvo estimación de las probabilidades de cada y intervalo y la sustitución de la χ^2_9 por la nueva χ^2_7, el contraste se realiza de exacta igual forma.

Siguiendo con los test del segundo capítulo nos encontramos con dos opciones. La primera consiste en aprovechar que nuestra hipótesis nula supone la normalidad de la muestra para aplicar el test de Lilliefors. Para ello estandarizamos nuestros datos empleando la media y cuasivarianza muestrales, y calculamos el estadístico de KS usando como función de distribución hipotética la de la normal estándar. Para este método contamos con tablas donde se recogen los valores críticos asociados a cada nivel de significación para cada n. El p-valor se calcula como: $0{,}886/\sqrt{n}$, para $\alpha = 0{,}05$ y cada valor de n.

La segunda posibilidad consistiría en realizar el mismo procedimiento pero empleando en vez los estadísticos de CvM y AD. En teoría esto sería posible debido a que como la normal es invariante a cambio de locación y escala, si se estandariza todos los estadísticos siguen un distribución fija (la de la normal estándar). El problema es que en la bibliografía seguida no se muestran tablas con sus p-valores más importantes, ni tampoco hemos conseguido una fuente fiable en internet. La solución que podemos tomar es emplear MonteCarlo para hallar el cuantil empírico de orden 95 y utilizarlo luego para realizar el contraste. Hemos obtenido 0,127 y 0,759 para CvM y AD respectivamente.

Finalmente, del tercer capítulo solo tenemos un test, el propuesto por Fan (1994). Su cálculo no tiene ninguna dificultad añadida a la primera sección una vez que tenemos los estimadores de la media y la varianza, después podemos emplearlos para especificar la función de densidad en el producto de convolución y hallar el valor del estadístico. También es necesario modificar ligeramente la fórmula de la varianza asintótica.

Ahora podemos intentar estudiar el error empírico de nuestros test y contrastar si nuestros estadísiticos se acercan a la distribución asintótica que les corresponde. Para los test del segundo capítulo no tenemos distribuciones asintóticas y dado que el punto crítico

	Test						
	$\hat{X}^2$	$\hat{G}^2$	$\hat{PD}$	Lilliefors	CvM	AD	T
Valor	2,54	2,46	2,50	0,526	0,0424	0,269	0,00115
Resultado contraste	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo

Cuadro 4.10: Tabla con los resultados de los test para la muestra

lo hemos simulado nosotros para los estadísticos de Anderson-Darling, solo tendría interés comprobar que la fórmula dada por Lilliefords efectivamente corresponde al cuantil de orden 95. De cualquier forma haremos el test para los tres, y así veremos que los cuantiles calculados por MonteCarlos son correctos.

Para el resto de test tenemos que estimar los parámetros, lo cual conlleva un proceso de optimización que es bastante costoso. Solo vamos a correr los test para $n = 100$ con mil repeticiones (tabla 4.11).

	Test						
	$\hat{X}^2$	$\hat{G}^2$	$\hat{PD}$	Lilliefors	CvM	AD	T
Error tipo I emp.	0,88	0,96[4]	0,91	0,954	0,946	0,947	0,980
p-valor del ks.test	$7{,}07e-6$	0,126	$7{,}8e-4$	No aplica	No aplica	No aplica	$8{,}88e-16$

Cuadro 4.11: Tabla con el error empírico y el resultado del test de KS para medir la adecuación de los estadístico a su comportamiento asintótico

Los resultados son quizás los esperables. Los estadísticos de divergencia parecen tender efectivamente a su comportamiento asintótico (el estadístico G^2 no ha estado definido para más de la mitad de las muestras, con lo que su valor está muy sesgado en perjuicio de aquellas que tienen intervalos vacíos), aunque se encuentren aún lejos de tal distribución. Probando a aumentar el valor de n parece que efectivamente su comportamiento se acerca cada vez más al de una χ^2_7, si bien el test que hemos diseñado tiene muchísima variabilidad, seguramente por causa del método aleatorio elegido para la elección de las fronteras. A falta de criterios más claros y de una mayor fuerza computacional, estos test, tal y como los hemos implementado, no parecen idóneos para el contraste de la hipótesis compuesta.

Los test de la segunda familia son los que tienen un comportamiento más regular. Obviamente, como nosotros mismos hemos estimado los cuantiles de CvM y AD, era esperable que el error empírico fuese muy cercano a 0,95. También se ha comprobado que la fórmula

[4]El estadístico no está definido para 696 muestras

de Lilliefors para hallar el p valor da una buena aproximación. Siguiendo estos resultados, parece natural que efectivamente estos test sean los más populares para medir la normalidad de entre todos los que hemos propuesto. Su desventaja es que están limitados a dicha distribución, o, adaptados, a otras distribuciones invariantes a cambios de escala y locación.

El test de Fan, como ya se podría anticipar a juzgar por la sección anterior, no tiene el comportamiento esperado. A pesar de que su error empírico no está demasiado lejos del esperable, su esperanza empírica es claramente negativa para todas las veces que hemos corrido el test, y su distribución está muy lejos de ser una normal estándar, dejando pocas esperanzas a que se normalice con una aumento de la n. La causa de este comportamiento errático probablemente se encuentre en el criterio de la elección de la h, que nosotros hemos establecido de forma bastante arbitraria (aunque cumpliendo con las restricciones asintóticas). Sería interesante entonces probar a establecer métodos de selección que tomen los anchos de banda que mejor se ajusten a la muestra, y estudiar su efecto en el comportamiento asintótico. Hemos probado con la regla del pulgar y el resultado no ha sido mejor, de hecho no suele ser recomendado en el contexto de los estimadores de densidad. El método que sí se suele sugerir es el de validación cruzada, pero precisa resolver un problema de optimización e implementarlo se haría muy pesado. Por tanto no hemos ido más allá y hemos decidido terminar aquí el análisis de nuestros test.

Conclusión

Empezábamos la introducción de este trabajo defendiendo la elección de las tres familias de test que hemos incluido sobre todos los tipos existentes de contrastes de especificación. Alegábamos que lo que las destaca por encima de las demás es su importancia histórica y conceptual, dos criterios que consideramos fundamentales para un trabajo que quiere funcionar como una introducción teórica a las técnicas más fundamentales con las que medir la divergencia entre una muestra y una distribución teórica.

Hemos visto, sin embargo, que su implementación en la práctica está lejos de ser trivial a pesar de lo intuitivo de su filosofía. Por supuesto, siempre se pueden emplear paquetes estadísticos que incluyan test ya diseñados que realicen el contraste en R o cualquier otro lenguaje de programación. Pero si intentamos adentrarnos en el código sobre el que se basan estos test, vemos que algunos de ellos, y en especial para la hipótesis simples, tienen que sortear problemas como la elección de los intervalos para el test de la X^2 de Pearson o el *bandwidth* para los estimadores de densidad kernel. Estas cuestiones que, en general, no tienen una respuesta convincente que surja de la teoría de la probabilidad, y que son abordadas utilizando técnicas computacionales. Intentar integrar en un solo papel la base teórica de cada una de las familias, junto con una descripción de los métodos computacionales que se usan para estudiarlas, e incluir los resultados que se han obtenido hasta la fecha sería una empresa tremenda. Este trabajo solo intenta ser una primera aproximación a ese ámbito mucho más grande que son los contrastes de especificación

A pesar de ello, hemos obtenido resultados bastante satisfactorios. Aún con lo rudimentario del código que hemos usado y de las limitaciones computacionales, hemos visto que para la hipótesis simple nuestros test tienen un buen comportamiento. Incluso, hemos obtenido resultados correctos cuando hemos aplicado los test del segundo capítulo al contraste de una hipótesis compuesta, y alentadores cuando usamos los estadísticos de divergencia. En todo caso, el método seguido dista de poder ser considerado riguroso, y siempre nos hemos limitado a la normal estándar y a una hipótesis nula correcta. Quedaría por ver, por ejemplo, qué ocurre cuando realizamos el contraste para una distribución discreta, o cuando nuestra hipótesis nula es diferente a la distribución que usamos para

simular nuestras muestras.

Ya hemos discutido los motivos por los cuales probablemente los estadísticos de Bickel y Rosenblatt no hayan rendido como era esperable. Lo cierto es que de entre las tres familias de contrastes, la basada en la densidad es la más reciente y menos estudiada. Además, la relativa complejidad de sus estadísticos hace que sea con diferencia la menos popular de las tres, a pesar del enorme interés que parece suscitar la estimación de la densidad en ámbitos como la econometría o las finanzas.

Sin embargo, la popularidad no ha sido un factor tan determinante a la hora de elegir los estadísticos. Si no, deberíamos haber hablado, por ejemplo, del estadístico de Shapiro-Wilk, que se encuentra entre los más populares para contrastar la normalidad de un conjunto de datos, y que se basa en la comparación de dos estimadores alternativos de la varianza de la muestra. Además, nos hemos dejado otras grandes familias de contrastes de especificación, que por lo menos querríamos introducir a continuación.

Cuando en el capítulo segundo introducíamos la fórmula general de un estadístico basado en la FDE usábamos la expresión:

$$T_n = c(n)d(\hat{F}_n, F_0),$$

con $d(.,.)$ una función de divergencia (aunque al final siempre hemos utilizado distancias).

Esta expresión es muy general, y nosotros siempre hemos hecho depender al estadístico de la FDE y la F_0 a través del proceso empírico. De esta forma teníamos que $B(x) = 0\ \forall x \Leftrightarrow F(x) = F_0(x)\ \forall x$, donde usábamos B para denotar el proceso empírico. Pero existen más maneras de caracterizar una distribución que a través de su función de distribución (o de su densidad, si la admite). También la función cuantil dada por $F^{-1}(x)$, o la función característica $\phi_F(x) = E_f[exp(ixY)]$ sirven para caracterizar una distribución. De esta forma, podríamos considerar otras opciones para B como:

$$B(x) = F^{-1}(x) - F_0^{-1}(x)$$
$$B(x) = \phi_F(x) - \phi_{F_0}(x)$$

Ambas posibilidades siguen cumpliendo la condición $B(x) = 0\ \forall x \Leftrightarrow F(x) = F_0(x)\ \forall x$, con lo que nos exponen dos otras grandes familias de contrastes de especificación: las basadas en el estimador de la función cuantil y el estimador de la función característica respectivamente. Abarcarlas en este trabajo era casi imposible por limitaciones de tiempo y espacio, con lo que nos conformamos con este pequeño esboce.

Aquí terminamos nuestro trabajo. A continuación presentamos el código que hemos usado en el capítulo cuarto. A pesar de ser rudimentario, y sin duda mejorable, los test de

los paquetes de R han confirmado que funciona correctamente. Lo hemos presentado tal y como ha quedado luego de cada análisis, comenzando con la definición de los parámetros del test de Monte Carlo, y terminando por sus resultados. En medio del cuerpo se genera la muestra y se calculan los estadísticos. Hemos incluido comentarios para que sea más legible.

Para más información sobre las diversas cuestiones que hemos tratado, o para profundizar más en los contrastes de especificación, recomendamos leer la bibliografía que presentamos al final del trabajo. En ella, destaca especialmente el manual *Comparing Distributions* de Oliver Thas, que ha servido como la principal referencia para este trabajo.

Código de R

.1. Hipótesis simple

.1.1. Primer capítulo

```
#DEFINIMOS EL TEST QUE NOS PERMITIRA REPETIR VARIAS VECES EL METODO
\begin{lstlisting}[language=R]
n=1000
rep=1000
tp=1:rep
tr=1:rep
tc=1:rep
vp=1:rep
vr=1:rep
vc=1:rep
for(j in 1:rep){
#GENERAMOS DATOS DE UNA NORMAL ESTANDAR
mu=0
sig=1
data=rnorm(n,mu,sig)

#Generamos intervalos equiprobables con prob 1/k
k=10
x=seq(0,1,1/k)
fron=qnorm(x,mu,sig)
p=1/k
vp=rep(p,10)

#Creamos la multinomial
```

```
mult=1:k
for(i in 1:k){
mult[i]=sum(fron[i+1]>data&data>fron[i])}

#CALCULAMOS EL ESTADISTICO DE PEARSON
x2=0
for(i in 1:k){
x2=x2+((mult[i]-n*p)**2)/(n*p)
}
#LO HE COMPROBADO USANDO chisq.test Y ES CORRECTO

#CALCULAMOS EL ESTADISTICO DE MAXIMA VER.
g2=0
for(i in 1:k){
g2=g2+2*mult[i]*(log(mult[i]/n)-log(p))}


#CALCULAMOS EL ESTADISTICO DE CRESSIE AND READ CON LAMBDA=2/3
lamb=2/3
pd=0
for(i in 1:k){
pd=pd+mult[i]*((mult[i]/(n*p))**lamb-1)}
pd=2/(lamb*(lamb+1))*pd
#HE COMPROBADO TOMANDO LAMBDA=1 QUE DA EL MISMO RESULTADO QUE X2

#RESULTADOS
#TEST PEARSON:
1-pchisq(x2,k-1)
#TEST RAZON DE VER.
1-pchisq(g2,k-1)
#TEST OPTIMO DE CRESSIE AND READ
1-pchisq(pd,k-1)

vp[j]=x2
vr[j]=g2
vc[j]=pd
```

```
tp[j]=1-pchisq(x2,k-1)>0.05
tr[j]=1-pchisq(g2,k-1)>0.05
tc[j]=1-pchisq(pd,k-1)>0.05
}

#RESULTADOS DEL TEST (EN OCASIONES G2 DA NaN CON LO QUE TENGO
#QUE IGNORAR LAS Na EN sum E INDICO CUANTOS HAY):
#TP:
sum(tp)/rep
#TR:
sum(tr,na.rm=TRUE)/(rep-sum(is.na(tr)))
sum(is.na(tr))
#TC:
sum(tc)/rep

#TESTEO QUE EFECTIVAMENTE LOS ESTADISTICOS SIGUEN UN CHISQ #USANDO EL ks.test
ks.test(vp,pchisq,k-1)
ks.test(vr,pchisq,k-1)
ks.test(vc,pchisq,k-1)
```

.1.2. Segundo capítulo

```
library("goftest")
#PARA PODER TRABAJAR CON LA DISTRIBUCION ASINTOTICA DE LOS ESTDISTICOS DE AD
#TENEMOS QUE UTILIZAR MONTECARLO
rep=10000
WA=rep(0,rep)
AA=rep(0,rep)

for(j in 1:rep){
for(i in 1:100){
WA[j]=WA[j]+(rnorm(1,0,1)/(i*pi))**2
}
}

for(j in 1:rep){
```

```
for(i in 1:100){
AA[j]=AA[j]+rnorm(1,0,1)**2/(i*(i+1))
}
}

#DEFINIMOS LAS DISTRIBUCIONES EMPIRICAS HALLADAS POR SIMULACION
FWA=ecdf(WA)
FAA=ecdf(AA)

#HALLAMOS LOS CUANTILES DE 95 ORDEN EMPIRICOS
AA=sort(AA)
WA=sort(WA)
qa=AA[9500]
qw=WA[9500]


#-------------------------------------------------------------------------------
#DEFINIMOS LOS PARAMETROS DEL TEST DE ANALISIS DEL ERROR DE LOS ESTADISTICOS
n=50
rep=1000

#VECTORES DONDE GUARDAREMOS LOS VALORES DE LOS ESTADISTICOS
vk=1:rep
vw=1:rep
va=1:rep

#VECTORES DONDE GUARDAREMOS LOS RESULTADOS DE LOS TEST
tk=1:rep
tw=1:rep
ta=1:rep

for(j in 1:rep){

#GENERAMOS DATOS DE UNA NORMAL ESTANDAR Y LOS ORDENAMOS
mu=0
sig=1
```

```
data=rnorm(n,mu,sig)
data=sort(data)

#CONSTRUIMOS LA FUNCION DE DISTRIBUCION EMPIRICA Y LA VISUALIZAMOS CON LA REAL
fde=ecdf(data)
x=seq(-4,4,0.01)
y=pnorm(x,0,1)
plot(fde)
lines(x,y,"l",lty=3,col="red")

#CONSTRUIMOS EL ESTADISTICO DE KS (SIN MULTIPLICARLO POR SQRT(N) DE MOMENTO)
D=0
for(i in 1:n){
dif1=abs(i/n-pnorm(data[i],mu,sig))
dif2=abs((i-1)/n-pnorm(data[i],mu,sig))
dif=max(dif1,dif2)
if(dif>D)D=dif
}

#DEFINIMOS LA FUNCION DE DISTRIBUCION ASINOTICA DEL ESTADISTICO (APROXIMADA PO
#LOS MIL PRIMEROS SUMANDOS):
FD<-function(x){
sum=0
for(i in 1:10000){
sum=sum+2*((-1)**(i+1))*exp(-2*(i**2)*(x**2))
}
val=1-sum
return(val)
}
#HALLAMOS EL P-VALOR:
1-FD(D*sqrt(n))

#PODEMOS COMPARARLO CON EL RESULTADO DEL ks.test() Y COMPROBAMOS QUE ESTA BIEN

#PROBAMOS AHORA CON LOS ESTADISTICOS DE ANDERSON-DARLING.
sum=0
```

```
for(i in 1:n){
sum=sum+(2*i-1)*(log(pnorm(data[i],0,1))+log(1-pnorm(data[n-i+1],0,1)))
}
A=-n-(1/n)*sum

W=0
for (i in 1:n){
W=W+(pnorm(data[i],0,1)-(2*i-1)/(2*n))**2
}
#PARA HALLAR EL RESULTADO DEL TEST
W<qw
A<qa

#HALLAMOS SUS P-VALORES
1-FWA(W)
1-FAA(A)

#GUARDAMOS LOS RESULTADOS
vk[j]=D
vw[j]=W
va[j]=A
tk[j]=1-FD(D*sqrt(n))>0.05
tw[j]=W<qw
ta[j]=A<qa
}
#RESULTADOS DEL MONTECARLO:
#ERROR TIPO I EMPIRICO
sum(tk)/rep
sum(tw)/rep
sum(ta)/rep

#ANALISIS POR KOLMOGOROV SMIRNOV
ks.test(sqrt(n)*vk,FD)
ks.test(vw,WA)
ks.test(va,AA)
```

.1.3. Tercer capítulo

```
library("kdensity")

#DEFINIMOS LOS PARAMETROS DEL ANALISIS DE LOS ESTADISTICOS
n=1000
rep=500
t1=rep(0,rep)
v1=rep(0,rep)
t2=rep(0,rep)
v2=rep(0,rep)
for(j in 1:rep){
#GENERAMOS LOS DATOS DE UNA NORMAL ESTANDAR
data=rnorm(n,0,1)

#CONSTRUIMOS EL HISTOGRAMA
hist(data)

#CONSTRUIMOS EL ESTIMADOR DE DENSIDAD KERNEL
x=seq(-5,5,0.01)
plot(x,dnorm(x,0,1),"l",lty=3,col="red")
h=5/sqrt(n)

kde=kdensity(data,kernel="gaussian",bw=h)
lines(kde)

#PRIMER ESTADISTICO DE BICKEL Y ROSENBLATT
int1<-function(x)(kde(x)-dnorm(x,0,1))**2
T1=integrate(int1,-Inf,Inf)$value

#SEGUNDO ESTADISTICO DE BICKEL Y ROSENBLATT
#TENEMOS QUE DEFINIR LA FUNCION RESULTANTE DEL PRODUCTO DE CONVOLUCION
con<-function(s) integrate(function(y, x=s) (dnorm((x-y)/h,0,1)/h)*dnorm(y,0,1)

#COMO con SE DEFINE CON EL COMANDO INTEGER, QUE NO ES UNA FUCCION VECTORIZADA,
#NO PODRIAMOS APLICAR DE NUEVO INTEGER PARA HALLAR EL VALOR DEL ESTADISTICO.
#SIN EMBARGO ESTE INCONVENIENTE PUEDE ARREGLARSE USANDO sapply Y DEFINIMOS ASI
```

```
#LA FUNCION: con VECTORIZADA
conv<-function(s) sapply(s,con)

#YA PODEMOS HALLAR FACILMENTE EL VALOR DEL ESTADISTICO
int2<-function(x) (kde(x)-conv(x))**2
T2=integrate(int2,-Inf,Inf)$value


#NOS INTERESA AHORA PODER REALIZAR EL TEST. TENEMOS QUE HALLAR LOS PARAMETROS.
#CALCULAMOS LA MEDIA Y LA VARIANZA ASINTOTICA
k2<-function(x) dnorm(x,0,1)**2
ik2=integrate(k2,-Inf,Inf)$value
mu=ik2/(n*h)

#PARA HALLAR LA VARIANZA TENGO QUE EVALUAR UNA INTEGRAL DOBLE.
#DEFINO LA FUNCION INTEGRANDO
kk<-function(u,v) dnorm(v+u,0,1)*dnorm(v,0,1)

#DEFINO LA FUNCION DE LA PRIMERA INTEGRAL (COMO ikk ES UNA INTEGRAL
#NO PUEDO APLICARLE integrate(), NECESITO VECTORIZARLA ANTES. DEFINO ikkv)
ikk<- function(s) integrate(function(v, u=s) dnorm(v+u,0,1)*dnorm(v,0,1), -Inf
ikkv<- function(s) sapply(s, ikk)

#LA ELEVO AL CUADRADO
ikkv2<-function(x) ikkv(x)**2

#YA PUEDO INTEGRARLA NORMALMENTE Y HALLAR LA VARIANZA DEL ESTADISTICO
sig=sqrt(2*integrate(ikkv2,-Inf,Inf)$value*ik2)

#PODEMOS YA REALIZAR LOS TEST EMPLEANDO LA DISTRIBUCION ASINTOTICA:
#TEST1:
1-pnorm(n*sqrt(h)*(T1-mu)/sig,0,1)
#TEST2:
1-pnorm(n*sqrt(h)*(T2-mu)/sig,0,1)
```

```
t1[j]=1-pnorm(n*sqrt(h)*(T1-mu)/sig,0,1)>0.05
t2[j]=1-pnorm(n*sqrt(h)*(T2-mu)/sig,0,1)>0.05
v1[j]=n*sqrt(h)*(T1-mu)/sig
v2[j]=n*sqrt(h)*(T2-mu)/sig
}

#RESULTADOS DE LOS ANaLISIS DE LOS ESTADISTICOS:
#ERRORES EMPIRICOS
sum(t1)/rep
sum(t2)/rep
#ks.test
ks.test(v1,pnorm,0,1)
ks.test(v2,pnorm,0,1)
```

.2. Hipótesis compuesta

.2.1. Primer capítulo

```
library(stats4)
rep=1000
tp=1:rep
tr=1:rep
tc=1:rep
vp=1:rep
vr=1:rep
vc=1:rep

for(j in 1:rep){
#GENERAMOS LA MUESTRA ALEATORIA
n=100
data=rnorm(n,0,1)
q=2

#CALCULAMOS EL ESTIMADOR MAXIMO VEROSIMIL DE LA MEDIA Y LA VARIANZA CON EL COM
#DEFINIMOS LA FUNCION LOG-VER:
lv<-function(mu,sig){
r=dnorm(data,mu,sig)
```

```
#
-sum(log(r))
}

#USAMOS mle PARA HALLAR LOS PARaMETROS QUE LA MINIMIZAN
mle(minuslogl=lv,start=list(mu=1,sig=1))

#OBTENEMOS NaNs DEBIDO A QUE EN EL PROCESO DE OPTIMIZACION LA FUNCION INTENTA
#ASIGNAR VALORES NEGATIVOS A sig. NO INFLUYE EN EL RESULTADO, LOS ESTIMADORES S
#CORRECTOS.
#ACCEDEMOS A ELLOS DE LA MANERA SIGUIENTE
mu=mle(minuslogl=lv,start=list(mu=1,sig=1))@coef[1]
sig=mle(minuslogl=lv,start=list(mu=1,sig=1))@coef[2]

f0<-function(x)dnorm(x,mu,sig)
F0<-function(x)pnorm(x,mu,sig)


#-----------------------------------------------------------------
#CAPiTULO 1
#GENERAMOS LA MULTINOMIAL
k=10
fron=1:k+1
fron[1]=-Inf
fron[k+1]=Inf
fron[2:k]=sort(runif(k-1,min(data),max(data)))

mult=1:k
for(i in 1:k){
mult[i]=sum(fron[i+1]>data&data>fron[i])
}

#CALCULAMOS LAS PROBABILIDADES ASOCIADAS A CADA INTERVALO
p=1:k
for(i in 1:k){
p[i]=integrate(f0,fron[i],fron[i+1])$value
```

```
}


#CALCULAMOS EL ESTADISTICO DE PEARSON
x2=0
for(i in 1:k){
x2=x2+((mult[i]-n*p[i])**2)/(n*p[i])
}

#CALCULAMOS EL ESTADISTICO DE MAX. VER.
g2=0
for(i in 1:k){
g2=g2+2*mult[i]*(log(mult[i]/n)-log(p[i]))}


#CALCULAMOS EL ESTADISTICO DE CRESSIE AND READ CON LAMBDA=2/3
lamb=2/3
pd=0
for(i in 1:k){
pd=pd+mult[i]*((mult[i]/(n*p[i]))**lamb-1)}
pd=2/(lamb*(lamb+1))*pd
#HE COMPROBADO TOMANDO LAMBDA=1 QUE DA EL MISMO RESULTADO QUE X2

vp[j]=x2
vr[j]=g2
vc[j]=pd

tp[j]=1-pchisq(x2,k-1-q)>0.05
tr[j]=1-pchisq(g2,k-1-q)>0.05
tc[j]=1-pchisq(pd,k-1-q)>0.05
}

#RESULTADOS DEL TEST:
#TP:
sum(tp)/rep
#TR:
```

```
sum(tr,na.rm=TRUE)/(rep-sum(is.na(tr)))
sum(is.na(tr))
#TC:
sum(tc)/rep


#TESTEO QUE EFECTIVAMENTE LOS ESTADISTICOS SIGUEN UN CHISQ
ks.test(vp,pchisq,k-1-q)
ks.test(vr,pchisq,k-1-q)
ks.test(vc,pchisq,k-1-q)
```

.2.2. Capítulo 2

```
library("goftest")
#COMO NO CONTAMOS CON TABLAS CON LA DISTRIBUCION DE AD Y CvM PARA TESTEAR NORM
#VAMOS A HALLAR EL CUANTIL DE ORDEN 95 POR MONTECARLO
rep2=10000
n=100
vw=1:rep2
va=1:rep

for(j in 1:rep2){
#GENERAMOS LOS DATOS. NOS VALE CUALQUIER TIPO DE NORMAL
data=rnorm(n,3,9)

#ESTANDARIZAMOS Y ORDENAMOS
data=(data-mean(data))/sd(data)
data=sort(data)

#HALLAMOS EL VALOR DE LOS ESTADISTICOS
sum=0
for(i in 1:n){
sum=sum+(2*i-1)*(log(pnorm(data[i],0,1))+log(1-pnorm(data[n-i+1],0,1)))
}
A=-n-(1/n)*sum
va[j]=A
```

```
W=0
for (i in 1:n){
W=W+(pnorm(data[i],0,1)-(2*i-1)/(2*n))**2
}
vw[j]=W
}

#HALLAMOS EL CUANTIL EMPIRICO
qam=sort(va)[rep2*0.95]
qwm=sort(vw)[rep2*0.95]
qam
qwm
#-----------------------------------------------------------------------
rep=1000
#VECTORES DONDE GUARDAREMOS LOS RESULTADOS DE LOS TEST
tk=1:rep
tw=1:rep
ta=1:rep
for(j in 1:rep){

#GENERAMOS DATOS DE UNA NORMAL ESTANDAR Y LOS ORDENAMOS
data=rnorm(n,0,1)
data=sort((data-mean(data))/sd(data))

#CONSTRUIMOS LA FUNCION DE DISTRIBUCION EMPIRICA Y LA VISUALIZAMOS CON LA REAL
fde=ecdf(data)
x=seq(-4,4,0.01)
y=pnorm(x,0,1)
plot(fde)
lines(x,y,"l",lty=3,col="red")

#CONSTRUIMOS EL ESTADISTICO DE KS (SIN MULTIPLICARLO POR SQRT(N) DE MOMENTO)
D=0
for(i in 1:n){
dif1=abs(i/n-pnorm(data[i],0,1))
```

```
dif2=abs((i-1)/n-pnorm(data[i],0,1))
dif=max(dif1,dif2)
if(dif>D)D=dif
}

#PROBAMOS AHORA CON LOS ESTADISTICOS DE ANDERSON-DARLING.
sum=0
for(i in 1:n){
sum=sum+(2*i-1)*(log(pnorm(data[i],0,1))+log(1-pnorm(data[n-i+1],0,1)))
}
A=-n-(1/n)*sum

W=0
for (i in 1:n){
W=W+(pnorm(data[i],0,1)-(2*i-1)/(2*n))**2
}
#GUARDAMOS LOS RESULTADOS
tk[j]=D<(0.886/sqrt(n))
tw[j]=W<qwm
ta[j]=A<qam
}
sum(tk)/rep
sum(tw)/rep
sum(ta)/rep

#LLAMAMOS A LOS PAQUETES QUE USAREMOS
library(kdensity)
library(stats4)

#DEFINIMOS LOS PARAMETROS DEL TEST
rep=100
n=500
v2=1:rep
t2=1:rep

for(j in 1:rep){
#GENERAMOS LA MUESTRA ALEATORIA
```

```
data=rnorm(n,0,1)
q=2

#CALCULAMOS EL ESTIMADOR MAXIMO VEROSIMIL DE LA MEDIA Y LA VARIANZA CON EL COM
#DEFINIMOS LA FUNCION LOG-VER:
lv<-function(mu,sig){
r=dnorm(data,mu,sig)
#
-sum(log(r))
}

#USAMOS mle PARA HALLAR LOS PARAMETROS QUE LA MINIMIZAN
mle(minuslogl=lv,start=list(mu=1,sig=1))

#OBTENEMOS NaNs DEBIDO A QUE EN EL PROCESO DE OPTIMIZACION LA FUNCION INTENTA
#ASIGNAR VALORES NEGATIVOS A sig. NO INFLUYE EN EL RESULTADO, LOS ESTIMADORES S
#CORRECTOS.
#ACCEDEMOS A ELLOS DE LA MANERA SIGUIENTE
mu=mle(minuslogl=lv,start=list(mu=1,sig=1))@coef[1]
sig=mle(minuslogl=lv,start=list(mu=1,sig=1))@coef[2]

f0<-function(x)dnorm(x,mu,sig)
F0<-function(x)pnorm(x,mu,sig)
#-----------------------------------------
#CONSTRUIMOS EL ESTIMADOR DE DENSIDAD KERNEL
h=1.06*sd(data)*n**(-1/5)

kde=kdensity(data,kernel="gaussian",bw=h)

#SEGUNDO ESTADISTICO DE BICKEL Y ROSENBLATT
#TENEMOS QUE DEFINIR LA FUNCION RESULTANTE DEL PRODUCTO DE CONVOLUCION
con<-function(s) integrate(function(y, x=s) (dnorm((x-y)/h,0,1)/h)*dnorm(y,mu,

#COMO con SE DEFINE CON EL COMANDO INTEGER, QUE NO ES UNA FUCCION VECTORIZADA,
#NO PODRIAMOS APLICAR DE NUEVO INTEGER PARA HALLAR EL VALOR DEL ESTADISTICO.
```

```
#SIN EMBARGO ESTE INCONVENIENTE PUEDE ARREGLARSE USANDO sapply Y DEFINIMOS ASI
#LA FUNCION: con VECTORIZADA
conv<-function(s) sapply(s,con)

#YA PODEMOS HALLAR FACILMENTE EL VALOR DEL ESTADISTICO
int2<-function(x) (kde(x)-conv(x))**2
T2=integrate(int2,-Inf,Inf)$value


#NOS INTERESA AHORA PODER REALIZAR EL TEST. TENEMOS QUE HALLAR LOS PARAMETROS.
#CALCULAMOS LA MEDIA Y LA VARIANZA ASINTOTICA
k2<-function(x) dnorm(x,0,1)**2
ik2=integrate(k2,-Inf,Inf)$value
mut=ik2/(n*h)

#PARA HALLAR LA VARIANZA TENGO QUE EVALUAR UNA INTEGRAL DOBLE.
#DEFINO LA FUNCION INTEGRANDO
kk<-function(u,v) dnorm(v+u,0,1)*dnorm(v,0,1)

#DEFINO LA FUNCION DE LA PRIMERA INTEGRAL (COMO ikk ES UNA INTEGRAL
#NO PUEDO APLICARLE integrate(), NECESITO VECTORIZARLA ANTES. DEFINO ikkv)
ikk<- function(s) integrate(function(v, u=s) dnorm(v+u,0,1)*dnorm(v,0,1), -Inf
ikkv<- function(s) sapply(s, ikk)

#LA ELEVO AL CUADRADO
ikkv2<-function(x) ikkv(x)**2

#HALLO LA INTEGRAL DEL CUADRADO DE LA FUNCION ESTIMADOR
ikd2=integrate(function(x) kde(x)**2,-Inf,Inf)$value

#YA PUEDO INTEGRARLA NORMALMENTE Y HALLAR LA VARIANZA DEL ESTADISTICO
sigt=sqrt(2*integrate(ikkv2,-Inf,Inf)$value*ikd2)

#PODEMOS YA REALIZAR EL TEST EMPLEANDO LA DISTRIBUCION ASINTOTICA:
#TEST2:
1-pnorm(n*sqrt(h)*(T2-mut)/sigt,0,1)
```

```
t2[j]=1-pnorm(n*sqrt(h)*(T2-mut)/sigt,0,1)>0.05
v2[j]=n*sqrt(h)*(T2-mut)/sigt
}

#RESULTADOS DEL TEST
sum(t2)/rep
ks.test(v2,pnorm,0,1)
```

Bibliografía

[1] Bickel, P.J. and Rosenblatt M. (1973). "On Some Global Measures of the Deviations of Density Function Estimate" . *The Annals of Statistics*

[2] Cressie, T.R.C. and Read, N.A.C., (1988) *Goodness of Fit Statistics for Discret Multivariate Data*

[3] Fan, Y., (1994) *Testing the goodness of fit of a parametric density function by kernel method*

[4] Gibbons, J.D. and Chakraborti, S. (1970) *Nonparametric Statistical Inference*

[5] Thas, O., *Comparing Distributions*

[6] Vélez, Ricardo, *Principios de inferencia estadística*

[7] Wand, M.P. and Jones, M.C., *Kernel Smoothing*